Non tutto ciò che può essere contato conta

e non tutto ciò che conta può essere contato

Albert Einstein

Indice

1 Processi INAR(1) **11**

1.1 Processo Autoregressivo del primo ordine 11

1.2 INAR(1) 13

1.3 Le distribuzioni della componente degli arrivi 15

1.3.1 Distribuzione Bernoulli 15

1.3.2 Distribuzione Binomiale 15

1.3.3 Distribuzione della Binomiale Negativa 16

1.3.4 Distribuzione Poisson 17

1.3.5 Distribuzione di Katz 18

2 Studio della dipendenza in modelli INAR(1) **21**

2.1 Score Test per $\alpha = 0$ 21

2.2 La statistica sotto l'ipotesi nulla 22

2.3 Quando gli arrivi sono processi parametrici 24

2.3.1 Binomiale 24

2.3.2 Poisson 25

2.3.3 Binomiale Negativa 26

2.4 Arrivi da Famiglie Parametriche 27

2.4.1 Famiglia Katz 28

2.4.2 Poisson Generalizzata 29

2.4.3 Teoria Asintotica 30

3 Esperimento di Monte Carlo **34**

4 Conclusioni **40**

A Approssimazione della Binomiale in una Poisson **42**

B Legge debole dei Grandi Numeri **43**

C Legge forte dei Grandi Numeri **44**

D Teorema del Limite Centrale per variabili I.I.D. **45**

E Teorema del Limite Centrale per variabili aleatore indipendenti **46**

Bibliografia **47**

Ringraziamenti

‘Desidero innanzittutto ringraziare la Professoressa Luisa Bisaglia per i suoi preziosi insegnamenti e per le numerose ore dedicate alla mia tesi. Inoltre, ringrazio i miei compagni di corso Francesco Gatti, Elisabetta Vablè,Elena Rogato, Valeria Brugnera, Ingres Artusi e Marco Serafin per tutto l’aiuto che mi hanno fornito in questo periodo. Infine, desidero ringraziare con affetto i miei genitori per il sostegno economico e per aver creduto nelle mie capacità,in particolare per essermi stati vicino nei momenti in cui credevo che questo giorno non sarebbe mai arrivato. ‘

Introduzione

Serie storiche per dati di conteggio, quindi a valori interi, possono incontrarsi in diversi contesti ed applicazioni, per esempio:

- il numero di autoveicoli che transitano nell'arco della giornata in un determinato tratto di strada;
- il numero degli aeri che in un giorno atterrano all'aereoporto di Malpensa;
- il numero di nuovi pazienti di cancro;
- il numero di reti segnate in una giornata del campionato di serie A di Calcio;
- il numero di nuovi 'disoccupati' nel mese di Gennaio;
- il numero d'insetti 'sopravvissuti' dopo l'uso dell'insetticida;
- il numero di nascite nel mese di Aprile;
- il numero di scorte presenti in magazzino durante l'anno commerciale;
- ecc.

Questi tipi di serie, per poter essere modellate in modo opportuno, richiedono l'utilizzo di modelli idonei che tengano conto della caratteristica di conteggio delle osservazioni. In letteratura sono

presenti diverse soluzioni. A tal proposito per una rassegna si può fare riferimento a Fokianos (2012).

In questa tesi viene trattato lo studio della dipendenza per le osservazioni di serie storiche per dati di conteggio attraverso l'utilizzo del modello INAR(1) (*first/order INtegervalued Auto-Regressive*) proposti da McKenzie (1985) e da Al-Osh e Alzaid (1987).

Jung e Tremayne in un lavoro del 2003 sostengono che quando si analizza una serie per dati di conteggio, prima di adattarla ad un modello INAR(1), bisognerebbe preliminarmente verificare l'ipotesi di indipendenza fra le osservazioni. In letteratura sono state proposte molte procedure per verificare la presenza di dipendenza, in una serie di dati di conteggio. In questo lavoro viene rivisitato un recentissimo test proposto da Sun e McCabe (2013) per la verifica della dipendenza seriale in dati di conteggio.

Questo lavoro estende quello sviluppato da Freeland (1998) e Jung e Tremayne (2003) attraverso una formula generale per una statistica *score* per verificare la dipendenza in un processo autoregressivo a valori interi quando la distribuzione degli arrivi è arbitraria. Infatti,la maggior parte dei lavori presenti in letteratura, considera la distribuzione di Poisson per modellare il processo degli arrivi. In particolare, vengono sviluppate due statistiche test capaci di tenere in giusta considerazione la sovra/sottodispersione (non solo l'equidispersione) del processo degli arrivi. Una caratteristica peculiare di come è formulato il modello è che il supporto dei processi dei dati in arrivo non è noto, cioè, non si conosce se sia finito od infinito,e se finito, non si conosce il limite superiore. Quindi quando si è in presenza di sottodispersione, può risultare conveniente utilizzare la distribuzione Binomiale che un supporto finito. Incorporando questa particolare caratteristica, il modello non assume una formulazione standard e il supporto della variabile può

dipendere da parametri ignoti.

In questo lavoro viene mostrato un test generale per l'indipendenza per un modello INAR con *thinning* binomiale con arrivi casuali su supporto sconosciuto. Quindi, se disponibile l'informazione a priori sulla natura del processo degli arrivi, questa può essere facilmente incorporata nel test. Tale concetto è approfondito nel paragrafo (2.3). Se le informazioni sullo stato di dispersione non sono disponibili, si possono utilizzare, per esempio le due famiglie parametriche di distribuzione quali la Famiglia di Katz o la Poisson Generalizzata. Sun e McCabe (2013) derivano la distribuzione della statistica test sotto l'ipotesi nulla e mostrano che il test è consistente per un grande insieme di processi generatori dei dati che presentano correlazione seriale di ordine uno.

Il presente lavoro è stato sviluppato in quattro capitoli:

- nel primo capitolo si richiama brevemente il modello INAR(1) e quali distribuzioni sono coinvolte nello studio;
- nel secondo capitolo viene descritto il test per la verifica dell'indipendenza nei modelli INAR(1) con le diverse distribuzioni degli arrivi;
- nel terzo capitolo attraverso uno studio di Monte Carlo, vengono studiati i comportamenti relativi al livello e alla potenza delle statistiche test con differenti processi di arrivo;
- nel quarto capitolo vengono illustrate le conclusioni di tale lavoro.

Capitolo 1

Processi INAR(1)

1.1 Processo Autoregressivo del primo ordine

Prima d'iniziare a parlare del modello INAR(1), mi sembra doveroso soffermarmi brevemente a spiegare come è fatto il modello autoregressivo del primo ordine, poichè esso è una parte integrante della formulazione del modello in esame.

Sia $\{\varepsilon_t\}$ un processo white noise di media nulla e varianza σ^2_ε. Si dice che $\{Y_t\}$ è un processo autoregressivo di ordine 1, se:

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t$$

Il processo sopra indicato lo abbiamo assunto senza la costante.

I suoi momenti primi e secondi sono:

Media $E\{Y_t\} = 0$.

Autocovarianza

$$\gamma_k = \begin{cases} \frac{\sigma_\varepsilon^2}{1-\phi_1^2} & k = 0 \\ \phi_1^k \gamma_0 & k > 0 \end{cases}$$

.

Operando per sostituzioni successive si ricava $\gamma_k = \phi^{k_1}\gamma_0$.

Dalla funzione di autococovarianza si ottiene la funzione di autocorrelazione (ACF): $\rho_k = \frac{\gamma_k}{\gamma_0} = \phi^{k_1} perk \neq 0$.

La funzione di autocorrelazione parziale è:

$$P_k = \begin{cases} \phi_1 & k = 1 \\ 0 & k > 1. \end{cases}$$

Si ricorda infine che un processo AR(1) è sempre invertibile e può essere scritto come un processo a media mobile di ordine infinito.

Per quanto riguarda la stazionarietà,si dimostra che un processo AR(1) è stazionario se le radici dell'equazione caratteristica sono, in modolulo maggiori di 1 .

Si può mostrare che un processo AR(1) ha l'ACF che tende ad annullarsi al divergere di k; in particolare a seconda del valore dei parametri ϕ_i, ρ_k tende a zero con un comportamento misto tra l'esponenziale e lo pseudoperiodico. Relativamente alla funzione di autocorrelazione parziale si può invece mostrare che, al divergere di k, essa è diversa da zero per $k \leq 1$, e si annula per $k > 1$.

1.2 INAR(1)

Nel trattare serie storiche per dati di conteggio, appare certamente apprezzabile identificare dei modelli, i quali prendono in considerazione osservazioni di tipo discreto. I modelli della classe INAR (INteger valued AutoRegressive) tengono conto di questa caratteristica delle serie di conteggio attraverso la definizione dell'operatore *thinning*.

La definizione dell'operatore *thinning* proviene da Steutal e Van Harn (1979). L'operatore *thinning* più comune è quello binomiale (denotato appunto dal simbolo $\circ$) ed è definito come segue.
Data una variabile aletoria discreta non negativa X_{t-1}, e un parametro $\alpha \in [0,1]$, l'operatore *thinning* è definito come

$$\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} B_{it} \tag{1.1}$$

dove $B_{1t}, B_{2t}, \ldots, B_{X_{t-1}t}$ sono successioni i.i.d. di variabili casuali Bernoulli con parametro α

$$P(B_{it} = 1) = 1 - P(B_{it} = 0) = \alpha \tag{1.2}$$

Dalla definizione dell'operatore $\circ$, è chiaro che: $0 \circ X_{t-1} = 0$, $1 \circ X_{t-1} = X_{t-1}$.

Si assume inoltre che B_{it} e ε_t siano indipendenti per tutti i e j.
Poichè $\alpha \circ X_{t-1}$; dato $X_{t-1} = x_{t-1}$ sono una somma di variabili casuali i.i.d. Bernulli allora per il Teorema del Limite Centrale si ha una distribuzione Binomiale con parametri α e x_{t-1}; $Bin(x_{t-1}, \alpha)$.

Una volta definito l'operatore *thinning*, il processo INAR(1) $\{X_t : t = 0, 1, 2, \ldots\}$ può essere definito come:

$$X_t = \alpha \circ X_{t-1} + \epsilon_t \tag{1.3}$$

con $\alpha \in [0, 1)$, $\{\epsilon_t\}$ è una successione i.i.d di variabili discrete non negative con media μ e varianza σ^2 e con un supporto $S_{\epsilon=\{0,1,2,\ldots,M\}}$ (M può essere finito od infinito; tipicamente il supporto di ϵ_t non è noto). Solitamente viene usata la distribuzione Poisson, infatti, essa è molto trattata in letteratura.

[1]

Quando siamo in presenza di $\alpha = 0$, segue che $S_{x_t} = S_\epsilon$. Il supporto di x_t dipende dal parametro sotto studio ed è dimostrato che la teoria classica della verosimiglianza, non è applicabile. Tuttavia, l'obiettivo dello studio è di derivare *one sided score test at* $\alpha = 0$, il quale ha una caratteristica aggiuntiva di essere il punto in cui viene delimitato lo spazio parametrico. Verrà assunto, come approccio, che le probabilità degli arrivi siano note, e di conseguenza si costruirà lo *score* test per α sotto questa assunzione.
Nel caso in cui le probabilità degli arrivi siano 'non note', esse verranno stimate sotto una varietà di assunzioni in modo tale che un test adeguato per d'indipendenza possa essere costruito con qualsiasi grado di adattabilità alla distribuzione degli arrivi. Così si può costruire un test d'indipendenza utilizzando le informazioni in modo che gli arrivi siano distribuiti come una Poisson o una Binomiale Negativa.

[1] La classe di modelli INAR(1) consente di ottenere una semplice e intuitiva interpretazione del processo generatore di dati; infatti, il valore di conteggio all'istante t può essere visto come la somma tra i sopravvisuti dal tempo *t-1* e i nuovi nati al tempo *t*. I sopravvissuti sono la realizzazione della variabile $\alpha \circ X_{t-1}$ e i nuovi nati sono la realizzazione dell'innovazione ϵ_t

Chiarito questo aspetto principale del modello, si può ora parlare delle famiglie parametriche coinvolte nel processo.

1.3 Le distribuzioni della componente degli arrivi

1.3.1 Distribuzione Bernoulli

Si consideri l'estrazione casuale di una unità da una popolazione dicotomica, cioè da una popolazione le cui unità sono raggruppate in due sole categorie, ad esempio 'sano' e 'non sano'. La singola esecuzione di tale esperimento va sotto il nome di *prova Bernoulliana*. Si associ il valore 1 all'evento A, designato, generalmente, con il termine 'successo', e il valore 0 all'evento B, il quale indica il termine 'insuccesso'. Sia p la probabilità di A e quindi 1-p la probabilità di B. Vale allora la seguente definizione di *variabile aletoria Bernulliana*.

Definizione 1.3.1. *Una variabile aletoria ha distribuzioone Bernoulliana se la sua funzione di probabilità è espressa da*

$$f(x) = p^x(1-p)^{1-x}$$

con x=0,1.

Tale distribuzione ha valore atteso pari a $E[X] = p$, e varianza $Var[X] = p(1-p)$.

1.3.2 Distribuzione Binomiale

La variabile aleatoria Binomiale è definita come il numero di successi in n prove Bernoulliane indipendenti.

Definizione 1.3.2. *Una variabile aletoria ha distribuzione binomiale se la sua funzione di probabilità è espressa da*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

dove x=0,1,. . . , n.

Il valore atteso è dato : $E[X] = np$ ciò significa che il valore atteso di successi che si verificano in n prove indipendenti quando la probabilità vale p, è pari a np. Siccome E[X]= np, otteniamo che

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= np[(n-1)p + 1] - (np)^2 \\ &= np(1-p) \end{aligned}$$

1.3.3 Distribuzione della Binomiale Negativa

Supponiamo di ripetere in maniera indipendente una prova, che abbia probabilità pari a p, $0 < p < 1$ di risultare in un successo, fintanto che non si totalizzano r successi. Se denotiamo con X il numero di prove necessarie per otternerli, allora

$$P[X = n] = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

con n=r, r+1, ...

Ha come valore atteso $E[X] = \frac{r}{p}$ e varianza $Var(X) = \frac{r(1-p)}{p^2}$.

1.3.4 Distribuzione Poisson

Il modello probabilistico che stiamo per introdurre è utile nelle applicazioni che riguardano fenomeni che si evolvono nel tempo e che implicano conteggi delle realizzazioni di un evento aleatorio. La variabile aleatoria di Poisson ha una vasta gamma di possibili applicazioni ad aree diverse, visto che può essere utilizzata come approssimazione di una variabile aleatoria binomiale di parametri (n,p), quando n è grande e p è piccolo a sufficienza perchè il prodotto np tende ad un valore positivo finito (la dimostrazione dell'approssimazione alla Binomiale verrà trattata nell'appendice). Diversamente dai casi precedenti, non si è in grado di introdurre la distribuzione Poisson facendo riferimento a meccanismi fisici (prove Bernoulliane, estrazioni casuali, ecc.); pertanto se ne dà immediatamente la definizione.

Definizione 1.3.3. *Si dice che una variabile aleatoria discreta ha distribuzione di Poisson con parametro* $\lambda > 0$ *se la sua funzione di probabilità assume la forma:*

$$f(X) = e^{-\lambda}\frac{\lambda^x}{x!}$$

con x= 0,1,2,....

La media e la varianza della distribuzione di Poisson sono date rispettivamente da: $E(X) = \lambda$, $Var(X) = \lambda$.

1.3.5 Distribuzione di Katz

Pearson (1895) ha osservato che per la distribuzione ipergeometrica [2]

$$\frac{p_x - p_{x-1}}{p_{x-1}} = \frac{a - x}{b_0 + b_1 x + b_2 x(x-1)} \tag{1.4}$$

dove a, b_0, b_1 e b_2 sono parametri, $p_x = Pr[X = x]$, e x assume valori interi positivi. Questo viene usato come punto di partenza per ottenere l'equazione differenziale che definisce il sistema di Pearson in distribuzioni continue. Pearson non ha continuato però, lo sviluppo per il caso discreto.

Nel 1919 Carver ha usato l'equazione alle differenze 1.4 per 'lisciare' dati 'attuariali'; anche se nel 1923 ottene delle espressioni per i termini parametrici dei momenti, non approfondì per le distriubuzioni discrete derivanti dalla 1.4.

Katz (1945) ha intrapreso uno studio per un caso speciale dove $b_0 = b_1$ e $b_2 = 0$.

Le restrizioni introdotte da Katz, e sopra indicate danno la *Katz Family distribuzione*, con

$$\frac{p_{x+1}}{p_x} = \frac{\alpha + \beta x}{1 + x} con x = 0, 1, \ldots, \tag{1.5}$$

dove $\alpha > 0$, $\beta < 1$ (se $\beta \geq 1$ non produce una distribuzione valida). Se $\alpha + \beta n < 0$, allora p_{n+j} è equavalente a 0 per tutti $j > 0$. Dalla 1.5,

$$(x+1)^{r+1} p_{x+1} = (\alpha + \beta x)(x+1)^r p_x \tag{1.6}$$

[2] La variabile aleatoria Ipergeometrica: supponiamo di dover estrarre un campione di n palline senza reinserimento da un'urna che contiene N palline, delle quali m sono bianche e $N - m$ sono nere. Se denotiamo con X il numero di palline bianche presenti tra le n estratte, allora $P(X = i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$ con $i = 0, 1, \ldots, n$ la variabile aleatoria ipergeometrica ha valore atteso pari a $E[X] = \frac{nm}{N}$ e varianza $Var[X] = \frac{N-n}{N-1} np(1-p)$ dove $p = \frac{m}{N}$

sommando entrambe le parti rispetto ad x, otteniamo:

$$\mu_{r+1}^{'} = \sum_{j=0}^{r} \binom{r}{j} (\alpha\mu_{j}^{'} + \beta\mu_{j+1}^{'}) \tag{1.7}$$

dove

$$\mu = \frac{\alpha}{1-\beta} \tag{1.8}$$

$$\mu_{2}^{'} = \frac{\alpha(\alpha+\beta)\mu}{\beta} \tag{1.9}$$

ovvero

$$\mu_2 = \frac{\alpha}{(1-\beta)^2} \tag{1.10}$$

Inoltre

$$\mu_3 = \mu_2(2c-1) \tag{1.11}$$

e

$$\mu_4 = 3\mu_2^2 + \mu_2(6c^2 - 6c + 1) \tag{1.12}$$

dove $c = \frac{\mu_2}{\mu} = (1-\beta)^{-1}$.

Katz ha mostrato che $\beta < 0$, $\beta = 0$, e $0 < \beta < 1$ (ovvero, $0 < c < 1$, $c = 1$, e $1 < c$) si dà luogo rispettivamente alla distribuzione Binomiale, Poisson e Binomiale negativa, con parametri $n = \frac{-\alpha}{\beta}$, $p = \frac{\beta}{\beta-1}$ per la binomiale, $\lambda = \alpha$ per la Poisson, e $k = \frac{\alpha}{\beta}$, $P = \beta$ per la binomiale negativa rispettivamente. Katz suggeriva inoltre, la riparametrizzazione per $\epsilon = \alpha/(1-\beta)$ $(= \mu)$ e $\eta = \beta/(1-\beta)$ $(= \left\{\sigma^2 - \mu\right\}/\mu)$ e ha dato le equazioni ML [3] (massima verosimiglianza) per le distribuzioni in termini di questi nuovi parametri.

Un importante motivazione per lo studio di Katz (1965), era il problema, una volta noto il data

[3] maximum likelihood

set di capire a quale di queste tre distribuzioni appartenesse.

Katz ha dimostrato che queste tre distribuzioni $Z = \frac{(s^2-\bar{x})}{\bar{x}}$, si approssimano ad una distribuzione Normale con media $(c-1)$ e varianza $2/N$ dove N è la dimensione del campione.

La funzione generatrice di probabilità della 'Katz Family' è

$$G(z) = \left(\frac{1-\beta z}{1-\beta}\right)^{-\alpha/\beta} \tag{1.13}$$

Capitolo 2

Studio della dipendenza in modelli INAR(1)

2.1 Score Test per $\alpha = 0$

Iniziamo l'analisi assumendo che $x_1, x_2, \ldots, x_t$ sia una serie di dati di conteggio generata dal seguente modello INAR(1)

$$X_t = \alpha \diamond X_{t-1} + \epsilon_t \tag{2.1}$$

e che gli arrivi del processo $\{\epsilon_t\}_{t=1}^{\infty}$ siano una serie di variabili casuali i.i.d. e con supporto S_ϵ che può essere sia infinito che finito.

Nell'approccio adottato nell'equazione (2.1) è sempre possibile testare la dipendenza dei dati di conteggio, utilizzando il coefficiente di autocorrelazione del primo ordine, che non dipende da nessuna delle ipotesi sul processo *thinning* o dalla distribuzione degli arrivi.

Date le equazioni (2.1) (1.2) e il vettore $(x_0, x_1, x_2, \ldots, x_t)$', la funzione di logverosimiglianza

(condizionata su x_0) è data data:

$$\begin{aligned}\mathcal{L} &= \sum_{t=1}^{T} \log[Pr(X_t = x_t)] \\ &= \sum_{t=1}^{T} \log\left[\sum_{k=0\vee\Delta x_t}^{x_t\wedge M} Pr(\alpha \circ X_{t-1} = x_t - K | X_{t-1} = x_{t-1}) p_k\right] \\ &= \sum_{t=1}^{T} \log\left[\sum_{k=0\vee\Delta x_t}^{x_t\wedge M} \binom{x_{t-1}}{x_t-K} \alpha^{x_t-K}(1-\alpha)^{(x_{t-1}-(x_t-K))} p_k\right]\end{aligned} \tag{2.2}$$

Nella quale $\Delta x_t = x_t - x_{t-1}$ e p_k è la probabilità associata agli arrivi del processo.

La generica espressione della derivata rispetto ad α è:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha} &= \sum_{t=1}^{T} \log\left\{\sum_{k=0\vee\Delta x_t}^{x_t\wedge M} \binom{x_{t-1}}{x_t-K} \alpha^{x_t-K}(1-\alpha)^{(x_{t-1}-(x_t-k))} p_k\right\} \\ &= \sum_{t=1}^{T} \frac{x_{t-1}}{1-\alpha}\left\{\frac{p(x_t - 1|x_{t-1} - 1)}{p(x_t|x_{t-1})}\right\}\end{aligned}$$

e si usa la convenzione che $p(x_t|x_{t-1}) = Pr(X_t = x_t|X_{t-1} = x_{t-1})$ e $p(i|j) = 0$ per entrambi $i < 0$ o $j < 0$.

2.2 La statistica sotto l'ipotesi nulla

Quando $\alpha = 0$ si ha$p(x_t|x_{t-1}) = Pr(X_t = x_t) = Pr(\epsilon_t = x_t) = p_{x_t}$. Quindi la derivata prima sotto l'ipotesi nulla è

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha} &= \sum_{t=1}^{T} x_{t-1}\left\{\frac{p_{x_t} - 1}{p_{x_t}} - 1\right\} \\ &= \sum_{t=1}^{T} x_{t-1} g(x_t) \\ &= S(p).\end{aligned} \tag{2.3}$$

Andiamo a porre $p_{-1} = 0$. Sotto l'ipotesi nulla di osservazioni i.i.d. abbiamo:

$$E[S(p)] = \sum_{t=1}^{T} E[x_{t-1}] E\left[\left\{\frac{p_{x_t} - 1}{p_{x_t}} - 1\right\}\right]$$

con x_t variabile casuale e con $S_\epsilon = 0, 1, 2, \ldots, M$ sotto l'ipotesi nulla

$$\begin{aligned} E\left[\frac{p_{x_t-1}}{p_{x_t}} - 1\right] &= \left[\sum_{k \in S_\epsilon} \left(\frac{p_{k-1}}{p_{pk}}\right) p_k\right] - 1 \\ &= \left[\sum_{k \in S_\epsilon} p_{k-1}\right] - 1 \\ &= -p_M = \mu_g \end{aligned} \tag{2.4}$$

e $p_\infty = 0 = \mu_g$ in questo caso dove il supporto è infinito. Così

$$\begin{aligned} E[S(p)] &= \sum_{t=1}^{T} E[x_{t-1}] \left\{\sum_{k \in S_\epsilon} p_{k-1} - 1\right\} \\ &= \begin{cases} -T\mu_x p_M & (M < \infty) \\ 0 & (M = \infty) \end{cases} \end{aligned} \tag{2.5}$$

pertanto, la derivata non ha media nulla. Se si dovesse utilizzare la statistica punteggio (2.3) quando la distribuzione degli arrivi possiede supporto finito 'non noto', (come vedremo nel caso della binomiale), ci porterebbe a non rifiutare mai l'ipotesi nulla per l'indipendenza per numerosità campionarie grandi.

Tipicamente gli *score* hanno media pari a zero ma sotto le condizioni standard di regolarità che qui non verranno applicate poichè il supporto x_t dipende dai parametri, come spiegato in precedenza. In aggiunta al fatto che lo *score* non ha media zero, la sua varianza non è data dalle usuali derivate seconde negative (Fisher Information).[1]. Ridefiniamo lo *score* con media μ_g

[1]Si dice infomazione attesa o Informazione di Fisher la quantità

$$i(\theta) = E_\theta(j(\theta)) = \left[-E_\theta\left(\frac{\partial^2 l(\theta)}{\partial\theta_r \partial\theta_s}\right)\right]$$

nell'equazione (2.4) ed otteniamo così:

$$S(p) = \sum_{t=2}^{T} x_{t-1}(g(x_t) - \mu_g) \tag{2.6}$$

la media di S(p) è zero a prescindere da ('non noto') M infinito o no. La varianza di questa statistica può essere calcolata direttamente per una qualunque assunzione sugli arrivi.

2.3 Quando gli arrivi sono processi parametrici

Nei modelli INAR, specialmente in quelli del primo ordine, nei processi di arrivo si usano le specificazioni delle distribuzioni marginali che hanno buone proprietà (per esempio: Poisson o Binomiale Negativa)

Tuttavia, se le informazioni sulle distribuzioni degli arrivi siano disponibili a priori, esse sono importanti e queste dovono essere utilizzate all'interno del modello. Qui di seguito verranno sviluppati test per diverse distribuzioni parametriche (specificatamente: Binomiale, Poisson e Binomiale Negativa) che fanno fronte a situazioni di sotto/equi e sovra- dispersione.
Nella pratica, parametri ignoti sono stimati sotto l'ipotesi nulla e le statistiche sono appropriatamente 'studentizzate' per garantine una distribuzione asintotica valida come vedremo in seguito.

2.3.1 Binomiale

Alcune caratteristiche rilevanti della Binomiale sono: quella di avere la media maggiore della varianza; un supporto finito.

---

valore atteso dell'informazione osservata

Nel caso dove ϵ_t è una binomiale, la funzione di probabilità è:

$$Pr(\epsilon_t = s) = (\binom{m}{s})p^s(1-p)^{m-s}, pers = 0, 1, \ldots, m \tag{2.7}$$

perciò

$$\frac{p_{x_t} - 1}{p_{x_t}} = \frac{x_t}{m - x_t - 1}\left(\frac{1-p}{p}\right) \tag{2.8}$$

poichè il supporto per ϵ_t è finito, allora $\mu_g = -p^m$ è diversa da zero e l'equazione (2.6) si riduce a:

$$S^{bin}(m, p) = \sum_{t=1}^{T} x_{t-1}\left(\frac{x_t}{m - x_t + 1}\left(\frac{1-p}{p}\right) - 1 - \mu_g\right) \tag{2.9}$$

Si ricorda infine che, la distribuzione Binomiale assume sottodispersione.

2.3.2 Poisson

Il modello PoINAR(1) presenta alcune limitazioni a causa dell'assunzione di equidispersione (media pari alla varianza del processo) e bassa flessibilità della distribuzione Poisson in quanto ha un solo parametro.

Nel caso in cui ϵ_t sia una Poisson , noi abbiamo

$$Pr(\epsilon_t = s) = \frac{e^{-\lambda}\lambda^s}{s!}$$

e

$$\frac{p_{x_t - 1}}{p_{x_t}} = \frac{x_t}{\lambda}$$

Il supporto di ϵ_t è infinito, allora $\mu_g = 0$ e la statistica è

$$S^p(\lambda) = \frac{\sum_{t=1}^{T} x_{t-1}(x_t - \lambda)}{\lambda} \tag{2.10}$$

2.3.3 Binomiale Negativa

La Binomiale negativa ha la caratteristica di permettere la sovradispersione, cioè consente di considerare una differente casistica di forme nella distribuzione del processo.
Come nel caso della Binomiale, la distribuzione marginale non è nota, però può essere calcolata considerando i primi due momenti, ma qui non verrà trattata.

Quando $\epsilon_t \backsim NBin(r, p)$ abbiamo

$$Pr(\epsilon_t = s) = \binom{s+r-1}{r-1} (1-p)^r p^s pers = 0, 1, 2, \ldots, \tag{2.11}$$

allora

$$\frac{p_{x_t-1}}{p_x t} = \frac{x_t}{r + x_t - 1} \frac{1}{p} \tag{2.12}$$

Il supporto per ϵ_t è inifinito, $\mu_g = 0$. Ricordiamo, come da convenzione $\frac{p_{x_t-1}}{p_{x_t}}$ è zero ogni volta che $x_t = 0$ ed allora, utilizzando la funzione indicatrice $I(\cdot)$ e l'equazione (2.6), otteniamo

$$S^{NBin}(r, p) = \sum_{t=1}^{T} x_{t-1} \left(I(x_t < 0) \frac{x_t}{r + x_t - 1} \frac{1}{p} - 1 \right) \tag{2.13}$$

Noi interpretiamo

$$I(x_t > 0) \frac{x_t}{r + x_t - 1}$$

$\frac{x_t}{r+x_t-1}$ è uguale a zero se ,x_t è zero, e questo evita la possibilità di una divisione non corretta

per r=1.

2.4 Arrivi da Famiglie Parametriche

In molte situazioni lo stato di dispersione dei dati non è noto a priori, e senza una qualche forma di *snooping* dei dati non è chiaro il test parametrico da utilizzare. Questa situazione può essere evitata considerando una famiglia parametrica di distribuzioni che può ospitare diversi gradi di dispersione. Un tale famiglia è quella di Katz (1965).

La Famiglia Katz è stata utilizzata per testare la dispersione di una serie di conteggi, grazie alla sua capacità di incorporare sotto , equi e sovradispersione. Katz (1965) ha sviluppato un test per l'ipotesi nulla della distribuzione equidispersiva Poisson contro l'ipotesi alternativa di sotto o sovradispersione per distribuzioni rispettivamente Binomiale e Binomiale Negativa.
Fang (2003) ha esteso il lavoro di Katz (1965) e ha sviluppato un test per discriminare le distribuzioni discrete nel contesto GMM.[2]
Yang (2009) ha inoltre utilizzato la famiglia Katz introducendo un test generale di sovradispersione. Tuttavia non c'è stato fino qui alcun lavoro che abbia incorporato la famiglia Katz nell'ambito dei modelli INAR.
Un'altra famiglia parametrica possibile è la Poisson Generalizzata, e questa permette la equi-,sovra- e sotto- dispersione. Tuttavia l'utilizzo della Poisson Generalizzata genera un problema di troncamento nel caso della sottodispersione e non ha una corretta distribuzione in quanto la somma delle probabilità non ammonta all'unità. Tuttavia, un test basato sulla Poisson Genera-

[2]Il metodo Generalizzato dei momenti, o GMM (dall'inglese Generalized Method of Moments), è stato sviluppato da Hansen (1982); estende il metodo dei momenti permettendo di trattare un numero di condizioni sui momenti maggiori del numero di parametri da stimare.
Gli stimatori GMM sono consistenti, asintoticamente normali ed efficienti nella classe di tutti gli stimatori.

lizzata sembra funzionare bene nella pratica, come si vedrà in seguito.

Nei seguenti paragrafi veranno illustate la Famiglia Katz e la Poisson Generalizzata.

2.4.1 Famiglia Katz

Una versione semplice, ma utile della famiglia della distribuzione Katz è definita dal rapporto delle probabilità

$$\frac{p_{j+1}}{p_j} = \frac{a+bj}{1+j} dovej = 0, 1, 2 \dots \tag{2.14}$$

$a > 0$, $b < 1$. Si noti che se j aumenta e $a + bj < 0$ allora metteremo tutte le succesive $p_{j+1} = 0$ per $i \geq 0$. Questo sistema ha casi particolari

- Bin(m,p) $b < 0$ $m = -a/b$ e $p = b/(b-1)$
- Poisson(λ) $b = 0$ $\lambda = a$
- Neg Bin(r,p) $0 < b < 1$ $r = a/b$ e $p = b$

Quando x_t si distribuisce come una Katz sotto l'ipotesi nulla, la media e la varianza sono

$$\mu_x = \frac{a}{1-b} \tag{2.15}$$

$$\sigma_x^2 = \frac{a}{(1-b)^2} \tag{2.16}$$

le quali si possono risolvere per a e b. Nonostante le generalità della famiglia Katz essa ha solo due parametri sconosciuti da calcolare. Richiamando la convenzione usata nell'equazione 2.13 è possibile ottenere

$$g^K(a, b, x_t) = I(x_t > 0)\frac{x_t}{(a-b)+bx_t} - 1 \tag{2.17}$$

e sfruttando l'equazione 2.6, osserviamo che

$$S^K(a,b) = \sum_{t=1}^{T} x_{t-1}\left(g^K(a,b,x_t) - \mu_g\right) \tag{2.18}$$

Si osservi che in questo caso non sappiamo se μ_g sia pari a zero.

2.4.2 Poisson Generalizzata

Supponiamo che ϵ_t segua la distribuzione Generalizzata di Poisson con parametri λ e κ, denotata da GP(λ,κ), allora ha funzione di probabilità

$$Pr(\epsilon_t = s) = \begin{cases} \lambda(\lambda+\kappa s)^{s-1}\exp^{-(\lambda+\kappa s)}/s! & s = 0,1,2,\ldots, \\ 0 & \text{per } s > m \text{ se } \kappa < 0. \end{cases}$$

dove $\lambda > 0$, $max(-1, -\lambda/m) < \kappa < 1$ e $m(\geq 4)$ è un numero intero che serve a soddisfare $\lambda + \kappa m > 0$, quando $\kappa < 0$. Una limitazione della Generalizzazione di Poisson è il troncamento dei dati di conteggio.

Normalmente la somma delle probabilità di variabili casuali discrete dovrebbe essere uguale all'unità, tuttavia per la distribuzione generalizzata di Poisson, la variabile casuale ϵ_t ha che $\sum_{s=0}^{m} Pr(\epsilon = s)$ è minore dell'unità quando $\kappa < 0$. Comunque in base a (Consul e Famoye, 2006), questo errore di troncamento è minore del 0.5% quando $m \geq 4$.

La distribuzione della Poisson Generalizzata si riduce a Pois(λ) se $\kappa = 0$. Quando x_t è generata sotto l'ipotesi nulla per la generalizzazione Poisson, la media e la varianza sono:

$$\mu_x = \frac{\lambda}{1-\kappa} \tag{2.19}$$

$$\sigma_x^2 = \frac{\lambda}{(1-\kappa)^3} \tag{2.20}$$

che possiamo risolvere per λ e κ.

Consentono di ottenere

$$g^{GP}(\lambda, \kappa, x_t) = \frac{[\lambda + \kappa(x_t - 1)]^{x_t - 2}}{(\lambda + \kappa x_t)^{x_t - 1}} e^{\kappa} x_t - 1 \tag{2.21}$$

e utilizzando l'equazione 2.6, abbiamo che

$$S^{GP}(\lambda, \kappa) = \sum_{t=1}^{T} x_{t-1} \left(\frac{[\lambda + \kappa(x_t - 1)]^{x_t - 2}}{(\lambda + \kappa x_t)^{x_t - 1}} e^{\kappa} x_t - 1 - \mu_g \right) \tag{2.22}$$

.

2.4.3 Teoria Asintotica

Si ha ora bisogno di una distribuzione asintotica per della statistica parametrica 'studentizzata' dell'equazione 2.6.

Viene definita

$$S(\theta) = \sum_{t=1}^{T} x_{t-1} \left(g(\theta, x_t) - \mu_g \right) \tag{2.23}$$

dove θ (px1) parametrizza la distribuzione p_{x_t}, ovvero

$$g(\theta, x_t) = \left\{ \frac{p_{x_{t-1}}(\theta)}{p_{x_t}(\theta)} - 1 \right\} \tag{2.24}$$

La media μ_g è definita come nell'equazione 2.4. Certamente i parametri devono essere stimati e verrà assunto che per $\hat{\theta}$ è disponibile che $T^{\frac{1}{2}(\hat{\theta} - \theta)}$ e $O_p(1)$ con $\mu_g = T^{-1} \sum_{t=1}^{T} g(\hat{\theta}, x_t)$. In aggiunta, la varianza di $S(\theta)$ deve essere stimata.

Teorema 2.4.1. *Assumiano che i dati* x_0, x_1, ...,x_T *siano i.i.d. con momento secondo finito. Si assume pure che* $T^{\frac{1}{1}}(\hat{\theta}-\theta)$ *è* $O_p(1)$ *e che* $g(\theta, x_t)$ *sia continua e differenziabile due volte per* θ*, dove* θ *è definita su un sottoinsieme compatto di* R^p*. Definiamo* $\hat{\mu_x}$ *e* s_x^2 *essere la media e la varianza delle* x_t *osservazioni, mentre* $\hat{\mu_g}$ *e* s_g^2 *sono la media e la varianza di* $g(\hat{\theta}, x_t)$*. Allora*

$$S(\hat{\theta}) = T^{\frac{-1}{2}} \sum_{t=1}^{T} \frac{(x_{t-1} - \hat{\mu_x})(g(\hat{\theta}, x_t) - \hat{\mu_g})}{s_x s_g} \xrightarrow{d} N(0,1) \tag{2.25}$$

Questo Teorema fornisce una distribuzione nulla, valida per $S(\hat{\theta})$.

Mentre $S(\hat{\theta})$ è progettato per avere rilevanza nel modello 1.3 con qualsiasi insieme specificato di arrivi è consistente (rigetta l'ipotesi nulla con probabilità 1) per un'ampio range di alternative. Qui di seguito scendiamo nel dettaglio. Se si è interessati a testare la correlazione in più situazioni generali; una correlazione negativa tra x_t e x_{t-1} è possibile e potrebbe essere auspicabile usare un test bilaterale; con correlazione positiva possono essere considerati e mantenuti test unilaterali. Se $\{x_t\}$ è assunto essere una sequenza di αmix, il quale include il modello 1.3, esso segue una sequenza di prodotti $\{y_t\} = \{(x_{t-1} - \mu_x)(g(\theta, x_t) - \mu_g)\}$ sono pure αmix della stessa grandezza. Quindi sotto l'assunzione di α mix, il quale ammette ad entrambi i processi, stazionari e non, la legge debole dei grandi numeri e il teorema del limite centrale di contenere $\{y_t\}$.

Per esempio, se $\{y_t\}$ è stazionario α mix,e con i momenti,

$$T^{\frac{-1}{2}} \sum_{t=1}^{T} [y_t - Ey_t] \xrightarrow{d} N(0, \omega^2)$$

dove $\omega^2 > 0$ è di solito la varianza di lungo periodo di $\{y_t\}$. Ma

$$S(\theta) = T^{\frac{-1}{2}} \sum_{t=1}^{T} (x_t - \mu_x)(g(\theta, x_t) - \mu_g) = T^{\frac{-1}{2}} \sum_{t=1}^{T} [y_t - Ey_t] + T^{\frac{1}{2}} Ey_y$$

così, quando $Ey_t \neq 0$,allora il test bilaterale basato su $S(\theta)$ sarà consistente e $S(\theta)$ diverge a $\pm\infty$. Segue che $S(\hat{\theta})$ è anche consistente eliminando gli effetti dei parametri stimati nello stesso modo come nel teorema sopra citato, ma sfrutta l'ipotesi del 'mixing'.

$Ey_t = E\left[(x_{t-1} - \mu_x)(g(\theta, x_t) - \mu_g)\right] \neq 0$ ogni volta che $Cov\left[x_t, x_{t-1}\right] \neq 0$ segue che $S(\hat{\theta})$ è consistente per ogni statzionario αmixing processo x_t con correlazione seriale del primo ordine. Segue lo stesso ragionamento per il test basato al ritardo 1 sia serialmente correlato e anche consistente. Come nel teorema limite dell'αmixing sotto la condizione di non stazionarietà, possiamo aspettarci la consistenza; verranno qui tralasciati i dettagli.

Il corollario che andremmo ad illustrare qui sotto è specifico nel caso di distribuzioni della famiglia Katz.

Corollario 2.4.1. *Assumiamo che i dati x_0,x_1,...,x_T siano i.i.d. con i momenti secondi finiti. Definiamo*

$$g^K(a, b, x_t) = I(x_t > 0)\frac{x_t}{(a-b) + bx_t} - 1$$

dove (a,b) si trovano in un sottoinsieme compatto di R^2. Le quantità a e b possono essere stimate

$$\hat{a} = \frac{\hat{\mu_x}}{s_x^2} \tag{2.26}$$

$$\hat{b} = 1 - \frac{\hat{\mu_x}}{s_x^2} \tag{2.27}$$

con $\hat{\mu_x}$ e s_x^2 essere la media e la varianza delle osservazioni x_t. Sono anche definite $\hat{\mu_g}$ e s_g^2 essere la media e la varianza di $g^k(\hat{a}, \hat{b}, x_t)$.

Allora

$$S^K(\hat{a}, \hat{b}) = T^{\frac{-1}{2}} \frac{\sum_{t=1}^{T}(x_{t-1} - \hat{\mu_x})(g^k(\hat{a}, \hat{b}, x_t) - \hat{\mu_g})}{s_x s_g} \xrightarrow{d} N(0, 1).$$

Allo stesso modo, abbiamo il seguente corollario per la generalizzazione della Poisson.

Corollario 2.4.2. *Assumiamo che i dati x_0,x_1,...,x_T siano i.i.d. con i momenti secondi finiti.*

Definiamo

$$g^{GP}(\lambda, \kappa, x_t) = \frac{[\lambda + \kappa(x_t - 1)]^{x_t - 2}}{(\lambda + \kappa x_t)^{x_t - 1}} e^{\kappa} x_t - 1$$

dove (λ, κ) appartengono ad un sottoinsieme compatto di R^2.

Le quantità λ e κ possono essere stimate

$$\hat{\lambda} = \sqrt{\frac{\hat{\mu_x^3}}{s_x^2}} \tag{2.28}$$

$$\hat{\kappa} = 1 - \sqrt{\frac{\hat{\mu_x}}{s_x^2}} \tag{2.29}$$

con $\hat{\mu_x}$ e s_x^2 essere la media e la varianza delle osservazioni x_t. Definiamo essere anche $\hat{\mu_g}$ e s_x^2 la media e la varianza di $g^{GP}(\hat{\lambda}, \hat{\kappa}, x_t)$.

Allora

$$S^{GP}(\hat{\lambda}, \hat{\kappa}) = T^{\frac{-1}{2}} \frac{\sum_{t=1}^{T} (x_{t-1} - \hat{\mu_x})(g^{GP}(\hat{\lambda}, \hat{\kappa}, x_t) - \hat{\mu_g})}{s_x s_g} \xrightarrow{d} N(0, 1).$$

Capitolo 3

Esperimento di Monte Carlo

Attraverso la simulazione Monte Carlo, ho potuto osservare come si comportano le statistiche 'Studentizzate' qui di seguito elencate:

Tabella 3.1: Statistiche test 'Studentizzate'

	Formule delle differenti Statistiche Test 'Studentizzate'
	Formule
Statistiche	Formule
Poisson	$S^P(\lambda) = T^{\frac{-1}{2}} \frac{\sum_{t=1}^{T}(x_{t-1}-\hat{\mu_x})(x_t-\hat{\mu_x})}{\hat{\mu_x}}$
Katz	$S^K(\hat{a},\hat{b}) = T^{\frac{-1}{2}} \frac{\sum_{t=1}^{T}(x_{t-1}-\hat{\mu_x})(g(\hat{a},\hat{b},x_t)-\hat{\mu_g})}{s_x s_g}$
Generalizzazione Poisson	$S^{GP}(\hat{\lambda},\hat{\kappa}) = T^{\frac{-1}{2}} \frac{\sum_{t=1}^{T}(x_{t-1}-\hat{\mu_x})(g(\hat{\lambda},\hat{\kappa},x_t)-\hat{\mu_g})}{s_x s_g}$
Autocorrelazione	$\hat{\rho} = T^{\frac{-1}{2}} \frac{\sum_{t=1}^{T}(x_{t-1}-\hat{\mu_x})(x_t-\hat{\mu_x})}{s_x^2}$

Andiamo ora a vedere nel dettaglio le statistiche sopracitate.

La statistica $S^P(\lambda)$ si basa sul modello 1.3 con un *thinning* binomiale e arrivi Poisson, ricavato da Freeland(1998), opportunamente 'studentizzata'.

Le statistiche $S^K(\hat{a},\hat{b})$ e $S^{GP}(\hat{\lambda},\hat{\kappa})$ sono state derivate sempre dal modello 1.3 con operatore *thinning* binomiale, ma, 'studentizzate' attraverso il Teorema (2.4.1.); mentre il coefficiente di

correlazione non richiede che siano specificati l'operatore 'thinning' e il processo di arrivo.
Come si può notare la differenza tra la Statistica Poisson e l'autocorrelazione è nel denominatore delle funzioni; e questo comporta che l'autocorrelazione avrà un test più potente se gli arrivi sono realizzazioni di una Poisson.

Prima di trattare le performance delle statistiche basate sull'operatore *thinning* binomiale nel modello INAR(1) con distribuzioni Poisson,Binomiale e Binomiale Negativa, diamo uno sguardo al caso in cui i processi degli arrivi siano discreti, come nelle misture di binomiali. Nella tabella seguente verranno illustrati i processi generatori dei dati:

Tabella 3.2: DGPs

	Possibili tipi di distribuzioni degli arrivi
	Formule
DGP	PFM
Pois(2)	$Pr(\epsilon_t = s) = \frac{e^{-\lambda}\lambda^s}{s!}$
Bin(5,0.4)	$Pr(\epsilon_t = s) = \binom{m}{s} p^s(1-p)^{m-s}$
Nbin(5,0.2857)	$Pr(\epsilon_t = s) = \binom{s+r-1}{r-1} p^s(1-p)^r$
$Bin(5,0.2) \wedge NBin(5,0.667)$	$Pr(\epsilon_t = s) = 0.5 * \binom{m}{s} p^s(1-p)^{m-s} + 0.5 * \binom{s+r-1}{r-1} p^s(1-p)^r$
$Bin(5,0.4) \wedge Bin(15,0.9)$	$Pr(\epsilon_t = s) = 0.5 * \binom{m1}{s} p_1^s(1-p_1)^{m-s} + 0.5 * \binom{m_2}{s} p_2^s(1-p_2)^{m-s}$

Per le simulazioni abbiamo assunto che sotto l'ipotesi nulla i processi generatori dei dati esposti, nella tabella 3.2 siano tutti indipendenti e identicamente distribuiti.
Abbiamo poi considerato due lunghezze delle serie T=100 e T=500, e abbiamo replicato ciascun modello per ciasuna grandezza 2000 volte.
Prima di parlare del nostro obiettivo primario e di queste simulazioni, è bene ricordare al lettore come è fatto il nostro modello in esame:

$$X_t = \alpha \circ X_{t-1} + \epsilon_t$$

Quello che siamo andati a verificare, con le simulazioni è la presenza/assenza di dipendenza dei dati in arrivo all'interno del nostro modello INAR(1), e abbiamo fissato l'ipotesi nulla che $\alpha = 0$, mentre l'ipotesi alternativa è $\alpha > 0$.

Come ci aspettavamo, la statistica $S^P(\hat{\lambda})$ è corretta solo in caso di equidispersione, mentre assume sotto o sovra dispersione dei dati, nei processi generatori Binomiale e Binomiale Negativa. Le statistiche test basate sulle famiglie di distribuzione Katz, Poisson Generalizzata e la correlazione sono delle buone statistiche e le useremo principalmente in quasi tutte le nostre simulazioni. Nella tabella 3.3, sono state calcolate le quattro statistiche per un modello INAR(1) con arrivi Poisson, e come si può notare la statistica Poisson è abbastanza equivalente all'autocorrelazione. Il punto cruciale a cui dobbiamo porre particolare attenzione per la nostra analisi è che i dati per un α^{oss} del 5% mi portano a rifiutare l'ipotesi nulla per un α superiore allo 0.03 nel caso in cui T sia 100; mentre se T è 500 rifuito l'ipotesi nulla per α superiori allo 0.01.

Quindi, all'aumentare di α parametro di *thinning* ci porta ad affermare che le osservazioni sono dipendenti.

Tabella 3.3: Statistiche Test sul modello INAR(1) - Pois(2)

INAR(1) -Pois(2)								
			α					
T=100	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$\hat{Stat.Pois}$	0.0260	0.0350	0.0625	0.0800	0.1125	0.1530	0.2185	0.2690
$\hat{Stat.Katz}$	0.0245	0.0355	0.0515	0.0765	0.1075	0.1540	0.2060	0.2690
$\hat{Stat.Gen.Pois}$	0.0285	0.0340	0.0470	0.0685	0.0960	0.1380	0.1915	0.2390
$\hat{\rho}$	0.000	0.0375	0.0555	0.0780	0.1055	0.1345	0.2055	0.2720
			α					
T=500	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$\hat{Stat.Pois}$	0.0395	0.0770	0.1930	0.3285	0.4860	0.6420	0.8060	0.8920
$\hat{Stat.Katz}$	0.0385	0.0755	0.1895	0.3275	0.4885	0.6520	0.8140	0.9000
$\hat{Stat.Gen.Pois}$	0.0390	0.0745	0.1845	0.3200	0.4595	0.6225	0.7935	0.8710
$\hat{\rho}$	0.0370	0.0755	0.1865	0.3295	0.4885	0.6515	0.8130	0.9025

Come possiamo vedere dalla tabella (3.4), le statistiche test basate sulle famiglie di distribu-

zione di Katz e Poisson Generalizzata tendono ad essere equivalenti. Per la verifica dell'indipendenza dei dati possiamo affermare che per $T = 100$, non si accetta l'ipotesi nulla per α superiori allo 0.05, mentre per T=500 rifiuto H_0 se $\alpha > 0.01$.
Anche in questo caso possiamo affermare che le osservazioni sono dipendenti.

Il modello sviluppato nella tabella 3.5 con arrivi Bernoulliani è un caso estremo di sottodispersione, e le statistiche domaninati sono quelle di Katz e Generalizzazione Poisson.
In questo caso si può notare facilmente che c'è dipendenza tra i dati ed il modello.

Tabella 3.4: Statistiche Test sul modello INAR(1) - Bin(5,0.4)

INAR(1) -Bin(5,0.4)								
			α					
T=100	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$\hat{Stat.Katz}$	0.0230	0.0350	0.0495	0.0820	0.1095	0.1430	0.2160	0.2550
$\hat{Stat.Gen.Pois}$	0.0235	0.0355	0.0495	0.0815	0.1105	0.1435	0.2160	0.2540
$\hat{\rho}$	0.0225	0.0340	0.0480	0.0795	0.1105	0.1420	0.2155	0.2510
			α					
T=500	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$\hat{Stat.Katz}$	0.0425	0.1050	0.2020	0.3520	0.5165	0.6710	0.7735	0.8735
$\hat{Stat.Gen.Pois}$	0.0380	0.0970	0.1940	0.3260	0.5000	0.6715	0.7760	0.8905
$\hat{\rho}$	0.0320	0.0885	0.1855	0.3235	0.4910	0.6695	0.7885	0.8910

Tabella 3.5: Statistiche Test sul modello INAR(1) - Bern(0.8)

INAR(1) -Bern(0.8)								
			α					
T=500	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$\hat{Stat.Katz}$	0.0515	0.2925	0.5685	0.7805	0.8885	0.9425	0.9590	0.9595
$\hat{Stat.Gen.Pois}$	0.0395	0.1090	0.2860	0.4990	0.7155	0.7735	0.8200	0.9345
$\hat{\rho}$	0.0350	0.0790	0.1790	0.3070	0.4810	0.6595	0.8060	0.9100

Nella tabella 3.6 si nota che quando gli arrivi sono distribuiti come una Binomiale Negativa, la statistica con una migliore prestazione è quella che è offerta dalla statistica $S^{GP}(\hat{\lambda}, \hat{\kappa})$.

Ricordiamo che solitamente la distribuzione Binomiale Negativa ha la caratteristica di sovradispersione dei dati,e osservando la tabella si nota che per $\alpha > 0.03$ si rifiuta l'ipotesi nulla per T=100, mentre per T=500 si rifiuta H_0 per $\alpha > 0.01$.

Tabella 3.6: Statistiche Test sul modello INAR(1) - NBin(5,0.2857)

INAR(1) -NBin(5,0.2857)								
			α					
T=100	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$Stat.\hat{K}atz$	0.0160	0.0275	0.0540	0.0960	0.1280	0.1660	0.2460	0.2995
$Stat.\hat{G}en.Pois$	0.0125	0.0225	0.0450	0.0755	0.1140	0.1435	0.2230	0.2730
$\hat{\rho}$	0.0200	0.0330	0.0565	0.0875	0.1180	0.1580	0.2240	0.2620
			α					
T=500	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$Stat.\hat{K}atz$	0.0395	0.0965	0.2155	0.3745	0.6045	0.7710	0.8935	0.9595
$Stat.\hat{G}en.Pois$	0.0235	0.0580	0.1365	0.2785	0.4710	0.6500	0.8260	0.9030
$\hat{\rho}$	0.0340	0.0815	0.1790	0.3200	0.4910	0.6580	0.8195	0.9080

In queste ultime due tabelle 3.7 3.8, vengono considerati modelli INAR con processi in arrivo che seguono distribuzioni non appartenenti alla Famiglia Katz. Abbiamo messo insieme in una prima analisi, la distribuzione binomiale e la distribuzione binomiale negativa, e come ultima analisi sono state combinate insieme due binomiali.

Nella tabella 3.7, per T=100 non riufiuto l'ipotesi nulla per $\alpha < 0.05$, mentre per T=500 rifiuto H_0 per $\alpha < 0.01$.

Nel Modello INAR(1) -Bin (5,0.4) e Bin(15,0.9) , notiamo che la statisca migliore è quella di Katz, per T=100 e che mi porta non riufiutare l'ipotesi nulla per $\alpha < 0.05$, mentre per T=500 rifiuto H_0 per $\alpha < 0.01$.

Tabella 3.7: Statistiche Test sul modello INAR(1) -Bin(5,0.2) e NBin(5,0.2857)

INAR(1) - Bin(5,0.2) e NBin(5,0.667)								
			α					
T=100	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$Stat.\hat{K}atz$	0.0205	0.0270	0.0485	0.0695	0.0980	0.1370	0.1715	0.2110
$Stat.G\hat{e}n.Pois$	0.0235	0.0290	0.0490	0.0765	0.1085	0.1560	0.1835	0.2330
$\hat{\rho}$	0.0235	0.0280	0.0520	0.0725	0.1120	0.1645	0.2065	0.2490
			α					
T=500	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$Stat.\hat{K}atz$	0.0355	0.0570	0.1095	0.1630	0.2555	0.3695	0.4845	0.5945
$Stat.G\hat{e}n.Pois$	0.0360	0.0720	0.1755	0.2645	0.4185	0.5740	0.7125	0.8355
$\hat{\rho}$	0.0385	0.0905	0.1925	0.3090	0.4885	0.6700	0.7960	0.8960

Tabella 3.8: Statistiche Test sul modello INAR(1) -Bin (5,0.4) e Bin(15,0.9)

INAR(1) -Bin (5,0.4) e Bin(15,0.9)								
			α					
T=100	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$Stat.\hat{K}atz$	0.0220	0.0365	0.0515	0.0675	0.1045	0.1575	0.2165	0.2730
$Stat.G\hat{e}n.Pois$	0.0200	0.0320	0.0425	0.0505	0.0770	0.1225	0.1710	0.2090
$\hat{\rho}$	0.0230	0.0400	0.0505	0.0665	0.1095	0.1575	0.2160	0.2775
			α					
T=500	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15
$Stat.\hat{K}atz$	0.0400	0.0895	0.1910	0.3300	0.4790	0.6595	0.8095	0.9100
$Stat.G\hat{e}n.Pois$	0.0320	0.0710	0.1455	0.2560	0.3855	0.5555	0.6925	0.8390
$\hat{\rho}$	0.0385	0.0835	0.1880	0.3375	0.4780	0.6710	0.8150	0.9100

Capitolo 4

Conclusioni

In questo lavoro viene illustrato come sia stata sviluppata una formula generale della Statistica *score* per lo studio della dipendenza in un modello INAR con casuali processi di arrivo e con diverse lunghezze dei ritardi.
Sono state prese in considerazione due statistiche nella quale venivano organizzati i processi degli arrivi che potevano essere sotto, equi e sovra dispersi. La prima statistica è quella basata sulla famiglia di distribuzioni Katz, ed includeva come casi speciali la Poisson, Binomiale e Binomiale Negativa. La seconda statistica test è quella basata sulla Poisson Generalizzata, ed essa include la distribuzione Poisson e può anche catturare per la sotto e sovradispersione.

Si è discusso della consistenza dei test sotto la distribuzione nulla; sono stati analizzati i livelli e le potenze dei test per diversi modelli, attraverso le simulazioni Monte Carlo.
Oltre alle statistiche sopracitate è stato preso in considerazione anche il coefficiente di autocorrelazione come punto di riferimento.

Dalle simulazioni Monte Carlo che sono state condotte, si è visto che la statistica Poisson non

è da ritenersi robusta, salvo il caso in cui gli arrivi siano anch'essi distribuiti come una Poisson. Le statitiche test basate sulle famiglie di Katz, Poisson Generalizzata e come il coefficiente di correlazione risultano essere robuste in tutti gli scenari considerati.

La statistica test basata sul coefficiente di correlazione si è dimostrata buona e generalmente competitiva, rispetto le statistiche test basate sulle famiglie di Katz e Poisson Generalizzata. In alcuni casi le statistiche test basate sulle famiglie di Katz e Poisson Generalizzata hanno mostrato una performance migliore rispetto alla statistica basata sul coefficiente di correlazione, come si è verificato per i modelli con distribuzioni degli arrivi miste e Bernoulliane.
Per le statistiche test basate sulle famiglie di Katz e Poisson Generalizzata si può tranquillamente affermare che sono le migliori indipendentemente dalla distribuzione dei dati d'arrivo.

Un'informazione importante per questo tipo di analisi sarebbe quella di conoscere a priori i processi generatori dei dati.

Gli esperimenti effettuati hanno evidenziato che, conoscere a priori come sono distribuiti i processi degli arrivi; permette di costruire test con buone proprietà, così da catturare la sotto, equi e sovradispersione.

Infine, dagli esperimenti è emerso che l'aumentare del valore di α parametro di *thinning*, porta correttamente a rifiutare l'ipotesi nulla e a concludere che le osservazioni di serie storiche per dati di conteggio sono dipendenti.

Appendice A

Approssimazione della Binomiale in una Poisson

Se eseguiamo una successione di n prove indipendenti, ognuna delle quali può risultare un successo con probabilità pari a p, allora, quando n è grande e p è sufficientemente piccolo da rendere np un valore positivo finito, il numero di successi che otteniamo vengono bene approssimati da una variabile aleatoria di Poisson con parametro $\lambda = np$. Questo valore di λ viene di solito determinato in maniera empirica.

Appendice B

Legge debole dei Grandi Numeri

Sia X_1, X_2, ..., una successione di variabili aleatorie i.i.d., ognuna con media finita $E[X_i] = \mu$. Allora, per ogni $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 quando n \to \infty$$

Appendice C

Legge forte dei Grandi Numeri

Sia X_1, X_2, ..., una successione di variabili aleatorie i.i.d., ognuna con media finita $E[X_i] = \mu$. Allora, con probabilità 1,

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \rightarrow \mu quandon \rightarrow \infty$$

Appendice D

Teorema del Limite Centrale per variabili I.I.D.

Sia X_1, X_2, ..., una successione di variabili aleatorie i.i.d, ognuna di media μ e varianza σ^2. Allora la distribuzione di

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

tende a una variabile aleatoria normale standard quando n tende ad infinito.

Ciò significa che, per $\infty < a < \infty$,

$$P\left\{\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{\frac{-x^2}{2}} dxquandon \to \infty$$

Appendice E

Teorema del Limite Centrale per variabili aleatore indipendenti

Sia X_1, X_2, ..., una successione di variabili aleatorie indipendenti, aventi rispettivamente, media $E[X_i] = \mu_i$ e varianza $V[X_i] = \sigma_i^2$. Se

- le X_i sono uniformemente limitate; cioè, se per un qualche M, $P_\infty \{|X_i| < M\} = 1$ per ogni i, e
- $\sum_{i=1}^{\infty} \sigma_i^2 = \infty$,

allora

$$P\left\{\frac{\sum_{i=1}^{n}(X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \leq a\right\} \rightarrow \Phi(a) quando n \rightarrow \infty$$

Bibliografia

[1] Jiajing Sun, Brendan P. McCabe *Score statistics for testing serial dependence in count data*, 1ª ed. online in Wiley Online Library ,(2012).

[2] R. Keith Freeland *Statistical Analysis of Discrete Time Series with application to the Analysis of workers' compensation claims data*,1ª ed. The University of British Columbia,(1998).

[3] M. A. AlOsh, A. A. Alzaid *First Order Integer Valued AutoRegressive(INAR(1)) Process*,1ª ed. Journal of Time Series Analysis,(1987).

[4] Robert. C. Jung, A. R. Tremayne *Testing for Serial Dependence in Time Series Models of Counts*,1ª ed. Eberhard Karls Universität Tübigen and University of Newcastle,(2001).

[5] L. Pace, A. Salvan *Introduzione alla Statistica*,1ª ed.Cedam,(2001).

[6] M. Ross *Calcolo delle Probabilità*,2ª ed. Apogeo,(2004).

[7] N. L. Johnson, A. N. Kemp, S. Kotz *Univariate Discrete distributions*,2ª ed. Wiley,(1992).

[8] G. Cicchitelli *Probabilità e Statistica*,2ª ed. Maggioli,(2004)