Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per l'economia e l'impresa

RELAZIONE FINALE

Regressione spuria fra Random Walks: un'analisi basata sulla concordanza/discordanza

Relatrice: Prof.ssa Luisa Bisaglia
Dipartimento di Scienze Statistiche

Laureanda: Chiara Varotto
Matricola: 1216762

Anno Accademico 2021/2022

Indice

Capitolo 1

Introduzione

Nello studio delle serie storiche, un assunto, che spesso viene ritenuto necessario e conveniente per le analisi, è quello per cui i dati osservati provengano da un processo stazionario. La maggior parte delle serie storiche economiche, tuttavia, presenta una componente tendenziale di fondo determinata principalmente dallo sviluppo e dall'evoluzione del sistema economico, che fa sì che media e varianza non risultino costanti nel tempo. Analizzando due di queste serie storiche, potrebbe verificarsi che, localmente, esse presentino trend stocastici comuni, che, di conseguenza, le portano a muoversi nella stessa direzione. Solo perchè due serie condividono uno stesso trend, non è detto che siano correlate tra loro, come potrebbe suggerire una prima analisi del coefficiente di correlazione di Pearson. Si parla in questo caso di regressione spuria o *nonsense regression.*

Questo elaborato si pone l'obiettivo di dimostrare, tramite simulazioni, come tra due *random walks* (indipendenti) di numerosità campionaria ristretta, sia frequente riscontrare andamenti casuali simili e, quindi, come la *nonsense correlation* rappresenti la regola piuttosto che l'eccezione. Per fare questo, vengono studiati periodi di tempo in cui le due serie storiche presentano associazioni concordi (tendono verso la stessa direzione) o discordi (tendono in direzioni opposte). Viene definita, in particolare, la massima lunghezza di istanti di tempo consecutivi di concordanza o discordanza e viene calcolata la probabilità di osservare valori uguali o maggiori della moda di questa lunghezza.

Questa relazione si sviluppa come segue. Viene da prima fatta una breve rassegna sull'evoluzione storica dei concetti di regressione spuria e *nonsense regression.* Viene quindi proposto un confronto rispetto al tema della cointegrazione. Nel capitolo 2 viene presentata la metodologia utilizzata per lo svolgimento di questo lavoro: uno

studio sulla concordanza e/o discordanza in *random walks* indipendenti, con una sezione dedicata alle distribuzioni approssimate. Le analisi numeriche a dimostrazione della teoria presentata, sono esposte nel capitolo 3. Nell'ultimo capitolo, viene invece considerata un'estensione del caso, rilasciando l'ipotesi di indipendenza e trattando *random walks* correlati.

1.1 Una breve rassegna storica

Il tema della regressione spuria è stato uno dei principali argomenti di studio a partire dalla fine del diciannovesimo secolo. Al centro dell'attenzione per la capacità di creare nuovi enigmi, piuttosto che risolverne, questa terminologia venne applicata a casi studio diversi, subendo quindi una continua variazione nel significato. Ogni studioso sembrava voler dare una propria interpretazione e soluzione al problema, ma è proprio grazie a tutti i dibattiti e confronti che si è potuta raggiungere una conoscenza più ampia del fenomeno.

Già solo il termine "correlazione" portava con sè numerose accezioni. Nelle prime pubblicazioni di Pearson (1892), si parlava di correlazione come debole forma di causalità. Per Galton (1888), invece, la correlazione tra due variabili rappresentava una misura di quanto esse fossero governate da *"common causes"*, quelle che Pearson definì qualche anno dopo come *"indipendent contributory causes"* (Pearson 1896). Già a partire dall'anno seguente però, quest'ultimo si rese conto di alcuni casi in cui non era presente alcuna forma di relazione organica e in cui, di conseguenza, la correlazione poteva essere ingannevole e non più basata su una pura forma di causa-effetto (Pearson 1897). L'autore, in seguito, attribuì la colpa di queste anomalie a una scorretta raccolta ed elaborazione dei dati, ma non seppe spiegarne i motivi e nemmeno proporre soluzioni.

1.1.1 Il contributo di Yule

Nel 1899, venne scoperto uno dei primi casi di correlazione spuria. Pearson, insieme a Lee e Bramley-Moore, sollevò la possibilità che i dati provenissero da un miscuglio tra distinte popolazioni internamente incorrelate (Pearson, Lee e Bramley-Moore 1899). Questo tema venne ripreso da Yule (1903), il quale si rese autore di una considerazione che lasciò il segno nella storia della statistica, soprattutto in ambito medico e sociale. Egli evidenziò la presenza di *«fallacies that may be caused by the mixing of distinct records»* e facendo questo, descrisse per la prima volta, quello che negli anni successivi prese il nome di Paradosso di Simpson (Simpson 1951), un altro

caso in cui è possibile giungere a conclusioni su relazioni errate seppur analizzando in modo statisticamente corretto i dati a disposizione.

Di notevole importanza furono poi gli studi condotti negli anni successivi. Mentre Pearson, Lee e Elderton (1910) pubblicavano un metodo per circumnavigare il problema della correlazione spuria, Yule riconsiderò i precedenti studi di Pearson (1897) a proposito della correlazione tra misure di organi di crostacei e, generalizzando il problema, distinse tre casi che potevano emergere nello studiarne le relazioni: (i) quando le cause di cui vogliamo comprenderne la natura influenzano direttamente le grandezze assolute; (ii) quando influenzano il tasso e (iii) quando non è possibile conoscere direttamente le modalità con cui operano queste cause. Per il primo caso, egli suggerì di correlare le grandezze assolute, per il secondo di correlare i tassi, mentre per il terzo caso non presentò una soluzione, ma dichiarò che di qualunque tipo fossero le relazioni (o la mancanza di relazioni) tra le variabili, era possibile che queste fossero prodotte dalle variabili stesse (Yule 1910). Quest'ultimo aspetto andava completamente contro le teorie elaborate a quei tempi e per questo, quando l'articolo venne pubblicato su *Biometrika* (1913), Pearson lo etichettò come *"a very curious blunder"*.

Come detto nell'introduzione sulla regressione spuria, sono molti i sinonimi che sono stati sviluppati. Negli scritti di Yule, per esempio, capita innumerevoli volte di imbattersi nell'espressione *"illusory association"*. Con questo termine, egli si riferisce alla forma di correlazione spuria che intendiamo ai giorni nostri, la stessa di cui tratta Simon (1954), ossia la correlazione generata da una dipendenza con una terza variabile. Quello che veramente era illusorio però, non era l'associazione in sè, ma bensì l'esistenza di una relazione causale. In questo si potrebbe creare un parallelismo con Pearson che nel termine *spurious correlation*, come spurio indica la relazione causale dedotta, o almeno così pensava Yule. Per Pearson, in realtà, la relazione causale non era qualcosa da desumere dalla correlazione, perchè secondo lui causalità *era* correlazione.

1.1.2 L'analisi di serie storiche

In contemporanea con altri studi, Pearson e Yule si dedicarono all'analisi di serie storiche aspettandosi di trovare gli stessi risultati che erano emersi dagli studi passati svolti su eventi a confronto. In particolare, la correlazione presente tra due serie, era più che altro originata dal tentativo di esprimere le stesse come somme di funzioni deterministiche del tempo ed errori. Agli studi di Yule, collaborò il suo amico R.H.

Hooker, il quale si concentrò prevalentemente sull'analisi di movimenti oscillatori minori all'interno di due curve (Hooker 1901). Secondo quest'ultimo, per studiare questi, era necessario concentrarsi su deviazioni dal trend e considerare quindi come variazione stocastica della serie, la deviazione dal valore precedente piuttosto che dalla media dell'intero periodo*. Qualche anno dopo, lo stesso Hooker (1905) propose un nuovo metodo di analisi: studiare la correlazione delle differenze prime delle serie, giungendo così a una maggiore comprensione delle dinamiche di cambiamenti brevi e repentini. In questo modo, era possibile, dunque, identificare la presenza di relazioni di breve periodo tra le due variabili; cosa che veniva, invece, mascherata se considerato il livello nell'intero periodo.
A proposito dell'aumento di correlazione tra movimenti oscillatori, Yule sostenne invece che le cause potessero attribuirsi a due aspetti: una relazione di causa-effetto, o al fatto che fossero entrambi funzioni del tempo o di una terza variabile (o gruppo di variabili), con la quale fosse presente una forma di dipendenza (Yule 1921).

1.1.3 Una nuova forma di 'non' correlazione

L'anno che segnò la svolta negli studi della regressione spuria fu il 1926, quando Yule per la prima volta parlò di *nonsense correlation*:

> But what one feels about such a correlation is, not that it must be interpretated in terms of some very indirect catena of causation, but that it has no meaning at all; that in non-technical terms it is simply a fluke, and if we had or could have experience of the two variables over a much longer period of time we could not find any appreciable correlation between them. But to argue like this is, in technical terms to imply that the observed correlation is only a fluctuation of sampling, whatever the ordinary formula, for the standard error may seem to imply. (Yule 1926, p. 4)

Rispetto alla distribuzione esatta della correlazione sotto l'ipotesi nulla di indipendenza definita da Student nel 1908 (Student 1908), Yule notò che essa poteva essere fuorviante per osservazioni tra loro correlate. Per analizzare, dunque, la natura di serie con valori correlati nel tempo, Yule suggerì un processo integrato di livello due I(2) («*conjuct series the differences of which are themselves conjuct series*») e scoprì che la correlazione risultante presentava una distribuzione di frequenza con forma a U. Di conseguenza, l'ipotesi nulla di assenza di relazione tra le due serie, veniva rigettata per qualunque livello di significatività in favore di una correlazione vicina

*Per approfondimenti, si veda (Hendry e Juselius 2000, p. 18)

in modulo all'unità. Attraverso l'operazione di differenziazione di serie costruite a partire da funzioni trigonometriche e termini casuali di disturbo, Yule (1927) contribuì allo sviluppo delle serie storiche dando origine ai modelli autoregressivi, AR(p) (vedi Klein 1997).

1.1.4 Granger e Newbold: una prima prova sperimentale

Le scoperte di Yule sull'esistenza di correlazioni prive di senso non furono molto incoraggianti per lo studio delle serie storiche economiche e per questo il suo lavoro rimase isolato fino al 1974, quando Granger e Newbold ripresero in mano il tema (Granger e Newbold 1974). I due Autori concentrarono i loro studi sulla regressione spuria originata da una inappropriata formulazione della struttura di autocorrelazione degli errori del modello di regressione. Non trattare con sufficiente attenzione l'autocorrelazione poteva, infatti, portare a tre principali conseguenze: (i) inefficienza delle stime dei coefficienti di regressione; (ii) non attendibilità delle previsioni basate sulla regressione e (iii) invalidità degli usuali test di significatività sui coefficienti di regressione. Essi si concentrarono maggiormente sul terzo punto.

Nelle prime simulazioni, Granger e Newbold (1974, p. 115) considerarono il classico modello di regressione lineare:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \,,$$

con Y_t e X_t generati da *random walks* indipendenti di lunghezza 50 e calcolarono, per 100 simulazioni, il valore della statistica test:

$$S = \frac{|\hat{\beta}_1|}{\hat{S.E.}(\hat{\beta}_1)}$$

usata per testare la significatività di β_1. Usando un tradizionale test t con un livello di significatività pari al 5%, gli Autori si resero conto che l'ipotesi nulla di coefficiente pari a zero veniva rifiutata (erroneamente) all'incirca 3 volte su 4, portando, di conseguenza, ad accettare la presenza di una relazione tra le due variabili. In particolare, la deviazione standard risultava sottostimata per un fattore di 5.6 e questo voleva dire che, invece che considerare un valore critico approssimativamente intorno a 2.0 bisognava usare un nuovo valore pari a 11.2 per attribuire un livello di significatività al test del 5%.

Nella simulazione successiva, vennero trattati processi *random walks* e processi ARIMA, sviluppati qualche anno prima da Box e Jenkins (1970). Gli Autori dimostrarono in modo più evidente il rifiuto dell'ipotesi nulla di mancanza di relazione:

per regressioni di *random walks* indipendenti, la percentuale di rifiuto era del 76% in presenza di un solo regressore, mentre saliva già al 93% quando i regressori erano 3. Per quanto concerne invece processi ARIMA(0,1,1), si otteneva rispettivamente un rifiuto del 64% e del 82% (vedi Tabella 2, Granger e Newbold 1974, p. 116).

A conclusione dei loro studi quindi, Granger e Newbold affermarono che lavorando con *random walks* o inserendo nella regressione variabili non necessarie, trovare relazioni spurie era la regola piuttosto che l'eccezione.

1.1.5 La teoria asintotica di Phillips

Granger e Newbold avevano dimostrato la facilità di raggiungere regressioni spurie, tuttavia non avevano eseguito nessuna analisi analitica per indagare le cause esatte di questa distorsione e fu Phillips a riprendere gli studi da questo punto, presentando la sua teoria asintotica (Phillips 1986). Egli dimostrò che nelle simulazioni da loro condotte, l'usuale test di significatività t, non possedeva una distribuzione standard, bensì divergeva al divergere della numerosità campionaria T, portando così ad un aumento della distorsione. Per T$\rightarrow \infty$, anche il coefficiente di determinazione campionaria R^2 tendeva ad infinito, convergendo ad una variabile casuale degenere.

Secondo Phillips, qualora si applicasse la teoria asintotica all'equazione di regressione $y_t = \hat{\alpha} + \hat{\beta} x_t + \hat{u}_t$ per $t = 1, ..., T$, con y_t e x_t generate da *random walks* indipendenti, le usuali statistiche $t_a = \frac{\hat{\alpha}}{s_{\hat{\alpha}}}$, $t_b = \frac{\hat{\beta}}{s_{\hat{\beta}}}$, usate per valutare la significatività dei coefficienti di regressione $\hat{\alpha}$ e $\hat{\beta}$, dovrebbero essere divise per la radice della numerosità campionaria al fine di ottenere un risultato corretto, anche se con distribuzione diversa dalla N(0,1) (vedi Phillips 1986).

In conclusione, per serie storiche non stazionarie, uno studio di regressione richiedeva tipicamente dei metodi asintotici e dei risultati diversi da quelli generalmente usati dalla teoria econometrica di quei tempi. Gli usuali test di significatività t non presentavano distribuzioni standard, bensì divergevano al crescere della numerosità campionaria e non era, quindi, possibile usare i convenzionali valori critici. Proprio non prestando attenzione a questi dettagli, era possibile riscontrare anomalie nelle analisi e false relazioni tra le variabili.

1.1.6 Gli studi moderni

Nel 2017, Ernst et al. pubblicarono un articolo che sembrò chiudere ogni dubbio sull'enigma della *nonsense correlation* (Ernst, Shepp e Wyner 2017). Nella loro introduzione, sottolinearono da subito la distinzione tra correlazione spuria e volatile:

la prima è osservabile quando due serie storiche sono entrambe dipendenti da una terza serie storica non considerata; la correlazione volatile, invece, è da intendersi in presenza di serie storiche indipendenti tra loro (senza alcun riferimento ad altre variabili) e viene considerata, dunque, come sinonimo di *nonsense regression*. In particolare, quest'ultima viene definita volatile perchè la sua distribuzione è fortemente dispersa e generalmente elevata in valore assoluto. A testimonianza di quanto detto, si veda l'istogramma sulla distribuzione del coefficiente di correlazione empirica (vedi Ernst, Shepp e Wyner 2017, p. 1791): il 95% dei valori rappresentati è contenuto nell'intervallo [-0.83 , 0.83] , di conseguenza ben distante da una distribuzione t Student.

Per risolvere l'enigma di Yule, Ernst e i colleghi calcolarono il valore del momento secondo del coefficiente di correlazione empirica tra due processi indipendenti di *Wiener*[†] e attraverso elaborazioni analitiche, ottennero un valore pari a 0.2405 , a prova del fatto che la distribuzione della correlazione risulta molto dispersa.

1.2 La cointegrazione

A partire dalla metà del ventesimo secolo, iniziò l'interesse verso l'analisi delle dinamiche di serie storiche e dei problemi generati dalla decomposizione delle stesse. Come abbiamo visto negli studi condotti da Yule, una delle tante cause per cui la correlazione tra due serie poteva aumentare, era la possibile dipendenza delle stesse dal tempo e quindi la presenza di un trend comune. Possiamo distinguere due tipi di trend: trend deterministico e trend stocastico. Il primo fa riferimento a una serie che evolve seguendo l'andamento lineare (o non) del trend ed è prevedibile; si tratta quindi di non stazionarietà in media. Nel secondo caso invece, la serie è influenzata da un trend casuale che genera *shock* dall'effetto permanente e porta la varianza del processo a essere infinita (non stazionarietà in varianza).
L'attenzione degli studiosi di quegli anni, dunque, venne rivolta all'analisi degli equilibri in serie storiche. Era necessario combinare, opportunamente, la dinamica di breve periodo con le proprietà di equilibrio di lungo periodo suggerite dalla teoria economica: le variazioni della variabile dipendente non sono determinate soltanto dalle variazioni della variabile indipendente, ma anche dall'entità del disequilibrio creatosi nel tempo precedente. Per analizzare nel dettaglio queste relazioni, Sargan (1964) sviluppò un modello di correzione dell'errore per cui la proporzione di disequilibrio di un periodo veniva corretta nel periodo successivo. Fu in questo elaborato,

[†]I processi di Wiener sono processi autocorrelati nel tempo. Sono integrali di puro rumore, quindi valori di differenti istanti temporali sono correlati.

che l'Autore distinse tra oscillazioni di breve periodo ed equilibrio di lungo periodo. Egli lavorò a un modello in grado di mettere assieme queste due dinamiche: con le differenze prime delle variabili era possibile cogliere variazioni di breve periodo; inserendo, invece, il valore ritardato di entrambe le variabili, era possibile valutare anche cambiamenti di livello di lungo periodo. Il tema venne ripreso e approfondito negli anni successivi da Hendry e altri studiosi. Questi capirono che se tra due serie, esisteva una relazione di equilibrio di lungo periodo, allora queste avrebbero avuto andamenti simili dovuti al condizionamento del sistema economico (Hendry, Davidson, Srba et al. 1977). Queste considerazioni e quelle di Sargan, non convinsero Granger che rifiutò l'idea che un processo non stazionario potesse diventare stazionario solo differenziandolo e dichiarò che variabili integrate dovevano rimanere tali. Tuttavia, cercando di confutare il ragionamento di Hendry, Granger (1981) dimostrò alla fine il contrario e gli diede ragione: una combinazione lineare di variabili non stazionarie in media poteva diventare stazionaria. Collegando, quindi, le sue teorie sulla regressione spuria con gli studi sull'equilibrio a lungo termine tramite il modello di correzione dell'errore di Sargan, Granger (1981) coniò il concetto di cointegrazione. Formalmente, la cointegrazione rappresenta una relazione di equilibrio di lungo termine fra serie storiche, che esiste solo se le variabili sono: (i) integrate dello stesso ordine e (ii) esiste una combinazione lineare delle stesse che è stazionaria. Diremo, quindi, che due serie y_t e x_t sono definite cointegrate se:

- sono $I(d)$, con $d \neq 0$ e uguale per entrambe,
- dato il vettore non nullo $a = (a_1, a_2)$, esiste una combinazione lineare delle stesse, tale che $a_1 y_t + a_2 x_t$ sia $I(0)$.

La teoria proposta da Granger però, non convinse la comunità internazionale degli econometrici perchè troppo semplice. Solo la collaborazione finale con Engle nel 1987 permise di produrre i primi test richiesti per la validazione del nuovo concetto. Attraverso il teorema di rappresentazione di Granger, gli Autori provarono che cointegrazione e meccanismo di correzione dell'errore erano due modi diversi per chiamare la stessa cosa: per ogni sistema cointegrato esiste una rappresentazione ECM; se esiste una rappresentazione ECM e le serie sono integrate, allora esse sono cointegrate (Engle e Granger 1987).

1.2.1 Il rapporto con la regressione spuria

La cointegrazione nasce quindi dal bisogno di comprendere se due variabili integrate, o con radice unitaria, siano realmente correlate oppure no. Nella sezione precedente,

abbiamo visto come questo concetto abbia rappresentato una svolta anche nei confronti della regressione spuria. Risulta importante però, sottolineare le differenze che intercorrono tra i due termini.

Cointegrazione e regressione spuria sono concetti che possono essere visti come due lati della stessa medaglia. La base di partenza è uguale per entrambe perchè si ha a che fare con serie storiche non stazionarie, le strade però poi si dividono. Nella regressione spuria non esiste una reale relazione tra i due fenomeni considerati, l'aumento di correlazione è dovuto solo a una forma di dipendenza di entrambi con una terza variabile non presa in considerazione. Nella cointegrazione, invece, esiste un'effettiva relazione genuina, nonostante la natura non stazionaria dei dati.

Quando si ha a che fare con serie di questo tipo, è quindi opportuno prestare le dovute precauzioni nelle analisi dei dati. Sicuramente una conoscenza dei fenomeni può aiutare a produrre delle considerazioni a priori sulle relazioni esistenti, ma è altrettanto necessario applicare test statistici per distinguere tra i due casi. Nei classici modelli lineari, si deve verificare che i residui siano *white noise*; nel caso di test per la cointegrazione invece, i residui devono essere essere testati per la stazionarietà.

Esistono diversi possibili test per la cointegrazione. Uno dei più generali è il test di Johansen (1988), basato sulla generalizzazione multivariata dei processi autoregressivi (VAR). Per rimanere, però, nel caso bivariato, prendiamo di seguito, come esempio, il test di Engle e Granger (1987). Seguendo questa procedura, come prima cosa, è necessario verificare, tramite opportuno test a radice unitaria, che entrambe le variabili prese in esame siano integrate dello stesso ordine. Una volta appurato questo, si definisce il modello di regressione lineare $y_t = \beta_0 + \beta_1 x_t + u_t$ e tramite il metodo dei minimi quadrati (OLS), si procede alla stima del coefficiente di cointegrazione $\hat{\beta}_1$. Questa regressione, applicata a variabili integrate, genera stimatori consistenti solo se le serie sono cointegrate. L'ultimo passo da verificare è che la combinazione lineare formata con questo coefficiente $\hat{\beta}_1$ sia stazionaria, ossia occorre eseguire un test di Dickey-Fuller aumentato (ADF) sui residui di regressione $\hat{u}_t = y_t - \hat{\beta}_1 x_t - \hat{\beta}_0$. In caso di rifiuto dell'ipotesi nulla di non stazionarietà, è possibile confermare la presenza di cointegrazione tra le due serie.

In questa sezione, sono stati affrontati brevemente gli aspetti principali legati al tema della cointegrazione. Non essendo, però, argomento principale di questo lavoro, per altri approfondimenti si rimanda alla letteratura presente[‡].

[‡](vedi Engle e Granger 1987)

Capitolo 2

Una nuova metodologia

Nel capitolo precedente sono stati brevemente rivisitati gli studi condotti nei secoli scorsi a proposito di regressione spuria e *nonsense regression*. Diversi sono stati gli spunti introdotti: un errore nella specificazione del modello di regressione, la mancata attenzione verso test di significatività distorti, la divergenza del coefficiente di determinazione e degli usuali test t e infine la dimostrazione della dispersione della distribuzione di frequenza del coefficiente di regressione. Tutto questo, però, non spiega perchè sia presente una correlazione priva di senso tra campioni piccoli di *random walks*.

In questo capitolo esploriamo quindi una nuova metodologia, proposta dagli Autori Hassler e Hosseinkouchack (2022), basata sul concetto di concordanza e discordanza e vediamo che all'interno di campioni piccoli di *random walks* indipendenti, si incorre frequentemente in associazioni spurie tra le due serie storiche.

2.1 Concordanza e discordanza

Si consideri un *random walk* bivariato $(X_i, Y_i)_{i=0,1,\ldots,n}$ con X_i e Y_i definiti come:

$$X_i = X_{i-1} + \varepsilon_i \quad Y_i = Y_{i-1} + \eta_i \quad \text{per} \quad i = 1, \ldots, n \tag{2.1}$$

dove (X_0, Y_0) sono due valori iniziali arbitrari. Sia $\Delta = (1 - \mathrm{B})$, con B operatore di ritardo, si definiscono, quindi, le differenze prime $(\Delta X_i, \Delta Y_i) = (\varepsilon_i, \eta_i)$, secondo le seguenti assunzioni:

Assunto 1. *Sia* $(\varepsilon_i, \eta_i)_{i=1,\dots,n}$ *una sequenza di variabili casuali continue, indipendenti e identicamente distribuite con*

$$\begin{aligned} p_\varepsilon &:= P(\varepsilon_i < 0), \quad P(\varepsilon_i > 0) = 1 - p_\varepsilon \\ p_\eta &:= P(\eta_i < 0), \quad P(\eta_i > 0) = 1 - p_\eta \end{aligned}$$

$p_\varepsilon, p_\eta \in (0,1)$. *In aggiunta,* ε_i, η_i *sono indipendenti e almeno una tra* p_ε *e* p_η *è uguale a 1/2.*

Si definiscono poi i concetti di concordanza e discordanza. Diremo che le variabili definite in 2.1 sono concordanti nell'i-esimo intervallo se X_i e Y_i si muovono nella stessa direzione; discordanti se si muovono, invece, nella direzione opposta. Di seguito viene data la definizione formale in termini della funzione segno:

$$\text{sign}(x,y) = \begin{cases} 1 & \text{se} \quad xy > 0 \\ 0 & \text{se} \quad xy = 0 \\ -1 & \text{se} \quad xy < 0 \end{cases}$$

Definizione 1. Per concordanza nell'i-esimo intervallo si intende $\text{sign}(\Delta X_i, \Delta Y_i) = 1$. Discordanza nell'$i$-esimo intervallo indica che $\text{sign}(\Delta X_i, \Delta Y_i) = -1$. *

Per rendere più immediata la scrittura, definiamo C_i l'indicatore di concordanza come segue:

$$C_i = \begin{cases} 0 & \text{se sign}(\Delta X_i, \Delta Y_i) = 1 \\ 1 & \text{se sign}(\Delta X_i, \Delta Y_i) = -1 \end{cases}, \quad i = 1, \dots, n \tag{2.2}$$

Dall'assunzione 1, ricaviamo che:

$$p := \text{P}(C_i = 0) = 1 - p_\varepsilon - p_\eta + 2p_\varepsilon p_\eta = \frac{1}{2} = \text{P}(C_i = 1).$$

Definiamo, ora, *zero run* una sequenza di zeri consecutivi in $(C_i)_{i=1,\dots,n}$ e Z_n la lunghezza della più estesa *zero run*, ossia la misura della più lunga sequenza di istanti temporali consecutivi in cui X_i e Y_i si muovono nella stessa direzione. La probabilità che $Z_n = k$, $\text{P}(Z_n = k)$, per una data numerosità n e con $k \in \mathbb{I}$, può essere espressa in funzione dei numeri di Fibonacci†.

*$\Delta X_i = 0$ e $\Delta Y_i = 0$ sono stati esclusi con probabilità 1 per assunzione.

†La definizione adottata si riferisce a quella proposta da Spickerman e Joyner (1984)

Definizione 2. Una sequenza di Fibonacci di ordine l , $(f_m^{(l)})_{m=1,2,\dots}$ per $l \in \{1, 2, \dots\}$ è definita come

$$f_m^{(l)} = \sum_{i=1}^{l} f_{m-i}^{(l)} \quad \text{per} \quad m > l,$$

con $f_m^{(l)} = 2^{m-1}$ per $m = 1, \dots, l$.

Nota 1. Ad esempio, nel caso in cui $l = 1$, si tratta di una sequenza di uno. Per $l = 2$, ne risulta la classica sequenza di Fibonacci. Per $l = 3$, parliamo invece di sequenza di Tribonacci.

Dalla definizione sopra riportata, possiamo ottenere la distribuzione di probabilità di Z_n come segue:

Proposizione 1. *Siano $(X_i, Y_i)_{i=0,1,\dots,n}$ definiti dalle equazioni 1, con le proprietà enunciate nell'assunto 1. Si può quindi ricavare la distribuzione:*

$$P(Z_n < k) = \frac{f_{n+1}^{(k)}}{2^n}, \quad 1 \le k \le n$$

Dimostrazione. Per la dimostrazione di questo risultato, si faccia riferimento all'appendice dell'elaborato di Hassler e Hosseinkouchack (vedi pp.193, 2022) □

Dalla proposizione 1, segue subito che:

$$\mathrm{P}(Z_n = k) = \frac{f_{n+1}^{(k+1)} - f_{n+1}^{(k)}}{2^n}, \quad 1 \le k \le n \tag{2.3}$$

dove $Z_n = 0$ corrisponde a una sequenza consecutiva di n volte 1 con probabilità $\mathrm{P}(Z_n = 0) = \frac{1}{2^n}$.

Fino ad ora, attraverso la definizione della *zero run*, abbiamo trattato una sequenza consecutiva di zeri, quindi il caso di concordanza. Inseriamo ora un'altra quantità che ci permetta di generalizzare il fenomeno in esame. Siamo interessati al numero massimo di istanti consecutivi dove X_i e Y_i siano concordanti o discordanti. Questo corrisponde alla lunghezza massima di una *zero run* o di una *uno run* in (C_i). Definiamo, quindi, S_n come la misura della sequenza più lunga di 0 o 1.

Corollario 1. *Sotto le assunzioni della Proposizione1, possiamo assumere che*

$$P(S_n < k) = P(Z_{n-1} < k - 1) = \frac{f_n^{k-1}}{2^{n-1}}, \quad 2 \le k \le n,$$

con $\quad P(S_n < 1) = 0.$

Dimostrazione. Hassler e Hosseinkouchack (2022, p. 194). □

Dal corollario 1, possiamo ricavare la probabilità

$$\mathrm{P}(S_n < k) = \frac{f_n^{(k)} - f_n^{k-1}}{2^{n-1}}, \quad 2 \leq k \leq n \tag{2.4}$$

Dai risultati 2.3 e 2.4, ponendo $\mathrm{P}(S_n = 1) = 2^{1-n} = \mathrm{P}(Z_{n-1} = 0)$ otteniamo che

$$\mathrm{P}(S_n = k) = \mathrm{P}(Z_{n-1} = k - 1), \quad k = 1, \ldots, n. \tag{2.5}$$

2.2 Risultati approssimati

Nella sezione precedente sono state specificate le principali distribuzioni necessarie per l'elaborazione di questa metodologia. Per le stesse quantità, stabiliamo ora le distribuzioni approssimate. [‡] Prima di procedere con la spiegazione, è necessario specificare qualche dettaglio di annotazione:

- $\lfloor x \rfloor$ indica la parte intera di $x \in \mathbb{R}$
- $\{x\} := x - \lfloor x \rfloor$ definisce la parte frazionaria di x
- log_b si riferisce al logaritmo in base b, mentre ln si riferisce al logaritmo naturale.

Proposizione 2. *Sotto le assunzioni della proposizione 1, si ricava che:*

$$P(Z_n - \lfloor \log_2 n \rfloor < z) = F_n(z) + o(1), k \in \mathbb{I}$$

dove $F_n(z) := \exp\left(-2^{-(z+1-\{\log_2 n\})}\right)$.

Dimostrazione. Si veda (Teorema 4, Földes 1975b) □

In accordo con quanto stabilito nell'equazione 2.5, per grandi n si può affermare che $\mathrm{P}(S_n = k) \approx \mathrm{P}(Z_n = k - 1)$, da cui segue che $S_n \approx Z_n + 1$. La distribuzione di Z_n può a sua volta essere approssimata troncando una distribuzione di Gambel per mezzo della funzione F_n. Definiamo V_n come una variabile di Gumbel con parametri $\{\log_2 n\} - 1$ e $1/\ln 2$ tale che

$$E(V_n) = \{\log_2 n\} - 1 + \frac{\gamma}{\ln 2} \qquad Var(V_n) = \frac{\pi^2}{6}\frac{1}{\ln^2 2},$$

[‡] Quanto segue è stato estratto dagli studi condotti da Földes (1975a), e Révész (1990) e riportati in Hassler e Hosseinkouchack (2022).

dove $\gamma \approx 0.5772$ è la costante di Eulero. In particolare, F_n per come è stata definita nella proposizione 2, equivale alla funzione di ripartizione di una variabile Gumbel, per cui $F_n(v) = \mathrm{P}(V_n \leq v)$, con $v \in \mathbb{R}$. La moda di V_n: $\mod(V_n) = \{\log_2 n\} - 1$[§]. Dalla proposizione 2 ricaviamo quindi che $Z_n - \lfloor \log_2 n \rfloor \approx \lfloor V_n \rfloor$ in modo tale che:

$$\mathrm{P}(Z_n - \lfloor \log_2 n \rfloor \leq z - 1) \approx \mathrm{P}(V_n \leq z) = \mathrm{P}(\lfloor V_n \rfloor \leq z - 1).$$

Da questo segue:

$$S_n \approx \lfloor V_n \rfloor + \lfloor \log_2 n \rfloor + 1\,. \tag{2.6}$$

Dato che

$$\mathrm{P}(\lfloor V_n \rfloor = z - 1) = \mathrm{P}(z - 1 \leq Vn \leq z) = \int_{z-1}^{z} f_n(v)\, dv,$$

e definita $-1 < \mod(V_n) < 0$, ricaviamo $\mod(\lfloor V_n \rfloor) = -1$. Questo risultato spiega il valore riportato, in seguito, nella tabella 3.3: $\mod(S_n) = \lfloor \log_2 n \rfloor$.
L'approssimazione in 2.6 deriva dalla proposizione 2; tuttavia tale distribuzione non può essere interpretata come una distribuzione asintotica: la variabile casuale V_n con funzione di distribuzione F_n non converge con n, perchè la sua parte frazionaria $0 \leq \lfloor log_2 n \rfloor < 1$ non lo fa. Analizziamo più nel dettaglio quanto affermato. Dalla proposizione 2 possiamo affermare che

$$\mathrm{P}(S_n \leq k) \approx \mathrm{P}(Z_n < k) \approx \exp\!\left(-2^{-(k+1-\log_2 n)}\right) = F_n(k - \lfloor \log_2 n \rfloor). \tag{2.7}$$

Dunque $\mathrm{P}(S_n = k)$ può essere approssimata da $P_n(k)$ definita come:

$$\mathrm{P}(S_n = k) \approx P_n(k) := F_n(k - \lfloor \log_2 n \rfloor) - F_n(k - 1 - \lfloor \log_2 n \rfloor). \tag{2.8}$$

Scriviamo ora n in funzione di $s = 0, 1, 2$ e $B = \{25, 30, 35\}$, $n = 2^s B$, tale che $\lfloor \log_2 n \rfloor = s + \lfloor \log_2 B \rfloor$ con $\{\log_2(2^s B)\} = \{\log_2 B\}$. Osserviamo che $P_n(\lfloor \log_2(2^s B) \rfloor)$ è costante per ogni s,

$$P_n(\lfloor \log_2(2^s B) \rfloor) = \exp\!\left(-2^{\{\log_2 B\}-1}\right) - \exp\!\left(-2^{\{log_2 B\}}\right).$$

Possiamo affermare, quindi, che:

$$\mathrm{P}(S_n > \lfloor \log_2(2^s B) \rfloor) \approx 1 - F_n(0) = 1 - \exp\!\left(-2^{\{\log_2 B\}-1}\right).$$

[§]La densità $f_n(v)$ è massimizzata nella $\mod(V_n)$

Dati i risultati enunciati in 2.6, possiamo definire

$$\mathrm{E}(S_n) = \sum_{k=1}^{n} k\mathrm{P}(S_n = k) = \sum_{k=0}^{n-1}(k+1)\mathrm{P}(Z_{n-1} = k) = \mathrm{E}(Z_{n-1}) + 1$$

e dal teorema 2 di Gordon, Schilling e Waterman (1986), ricaviamo:

$$\mathrm{E}(Z_n) \approx \log_2 n + \frac{\gamma}{\ln 2} - \frac{3}{2}$$

da cui segue che:

$$\mathrm{E}(S_n) \approx \mu_n := \log_2 n + \frac{\gamma}{\ln 2} - \frac{1}{2}. \tag{2.9}$$

In particolare, Guibas e Odlyzko (1980, Teorema 4.1) dimostrarono che

$$\mathrm{E}(S_n) = \mu_n + r(n) + o(1)$$

dove $r(n)$ non tende a zero, ma rappresenta una quantità molto piccola $|r(x)| \leq 1.6 \times 10^{-6}$ per tutti gli x definiti in Guibas e Odlyzko (1980, p. 245). Dato che $r(n)$ non si annulla, S_n non converge con n, nonostante μ_n.

Gordon, Schilling e Waterman (1986, Teorema 2) hanno definito anche l'approssimazione per la varianza. Dato che $\mathrm{Var}(S_n) = \mathrm{Var}(Z_n)$, si ottiene[¶]

$$\mathrm{Var}(S_n) \approx \frac{\pi^2}{6\ln^2 2} + \frac{1}{12} \approx 3.5070\,.$$

[¶](Guibas e Odlyzko 1980, vedi Teorema 4.1)

Capitolo 3

Analisi numeriche

In questo capitolo vengono svolte delle analisi numeriche al fine di sostenere matematicamente e graficamente quanto dimostrato nei capitoli precedenti.

Consideriamo un *random walk* bivariato $(X_i, Y_i)_{i=0,1,\dots,n}$ definito come nell'equazione 2.1. Per le seguenti simulazioni, calcoliamo le differenze $(\Delta X_i, \Delta Y_i) = (\varepsilon_i, \eta_i)$ a partire da una distribuzione normale bivariata tale che:

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{3.1}$$

I valori di partenza (X_0, Y_0) vengono fissati pari a $(0, 0)$ e la correlazione campionaria viene calcolata come:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Sapendo che $\hat{\rho}$ è simmetrica rispetto allo zero per $\rho = 0$, prendiamo il valore assoluto $|\hat{\rho}|$.
Nella tabella 3.1, sono riportate le medie di 10^5 replicazioni dei valori assoluti di $\hat{\rho}$ di *random walks* con numerosità campionaria $n = \{25, 50, 100, 200\}$. Si ponga l'attenzione in particolare alla prima riga della tabella: per $\rho = 0$ si ha una forte evidenza a favore della correlazione spuria. Si può notare come i risultati rimangano all'incirca uguali al cambiare della numerosità e come il risultato sia notevolmente diverso dal valore di correlazione teorico. Per valori maggiori, ma non elevati, di ρ, si può ancora osservare come la correlazione campionaria sovrastimi i valori reali, mentre per $\rho = 0.6$ e $\rho = 0.8$ si raggiungono valori ottimali con $|\hat{\rho}| \approx 0.6$ e $|\hat{\rho}| \approx 0.77$.

Tabella 3.1: Media di 10^5 replicazioni di valori assoluti della correlazione campionaria di *random walks* con valori iniziali pari a 0.

n	25	50	100	200
$\rho = 0$	0.4255	0.4222	0.4217	0.4221
$\rho = 0.2$	0.4456	0.4421	0.4405	0.4414
$\rho = 0.4$	0.5029	0.5005	0.4992	0.4996
$\rho = 0.6$	0.6055	0.6044	0.6031	0.6037
$\rho = 0.8$	0.7652	0.7651	0.7640	0.7651

Nel prosieguo di questo capitolo, attraverso valutazioni numeriche, verrà calcolata la lunghezza massima di associazioni casuali tra *random walks* indipendenti, a partire da campioni di numerosità piccola e media, a spiegazione dell'esistenza di una sovrastima della correlazione campionaria rispetto a quella teorica, come si può evincere dalla tabella 3.1.

3.1 La quantità S_n

Di nostro interesse, è studiare la lunghezza massima di una *zero run* o di una *uno run*, quindi per le prossime analisi useremo la quantità S_n in modo da includere sia concordanza che discordanza.

Sulla base delle relazioni 2.4 e 2.5, nella tabella 3.2 sono state calcolate, tramite simulazioni, le misure statistiche principali relative alla quantità S_n. In particolar modo, possiamo osservare come il valore atteso aumenti di una unità al raddoppiare di n e questo ci porta ad affermare che abbia un andamento logaritmico. Allo stesso modo anche la varianza aumenta, anche se in modo meno significativo, mentre asimmetria e curtosi diminuiscono all'aumentare di n. Questi ultimi due aspetti sono messi maggiormente in rilievo nella figura 3.1.

Tabella 3.2: Valore atteso, varianza, asimmetria e curtosi calcolate per S_n sulla base delle probabilità del corollario 1.

	$n = 25$	$n = 50$	$n = 100$	$n = 200$
media	4.9799	5.9783	6.9774	7.9770
varianza	2.6419	2.9983	3.2134	3.3401
asimmetria	1.2601	1.2173	1.1759	1.1465
curtosi	5.7909	5.6780	5.5361	5.4366

Dalle analisi numeriche svolte su S_n, sono stati tratti quattro grafici mostrati nella figura sottostante. Questi rappresentano la distribuzione di frequenza di S_n al variare della numerosità campionaria, per valori di $k = 1, \ldots, 15$. Possiamo osservare come la probabilità massima diminuisca all'aumentare della numerosità campionaria: per $n = 25$, otteniamo una probabilità massima di 0.2980, per $n = 50$ di 0.2768, per $n = 100$ di 0.2645, per $n = 200$ di 0.2574 .

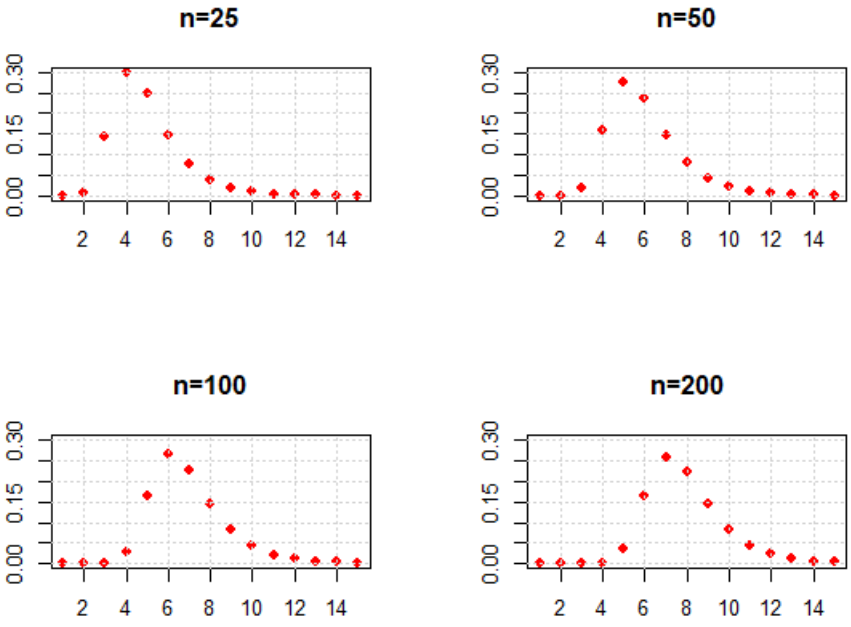

Figura 3.1: $\mathrm{P}(S_n = k)\,, k = 1, \ldots, 15$

Osservando meglio i grafici, possiamo notare inoltre che i 5 valori con la probabilità più alta includono all'incirca il 90% della probabilità totale:

$$\mathrm{P}(3 \leq S_{25} \leq 7) = 0.9195 \qquad \mathrm{P}(4 \leq S_{50} \leq 8) = 0.8995$$
$$\mathrm{P}(5 \leq S_{100} \leq 9) = 0.8850 \qquad \mathrm{P}(6 \leq S_{200} \leq 10) = 0.8758\,.$$

Si consideri, ora, come quantità di riferimento, per il calcolo delle probabilità, la moda di S_n. Dalle analisi numeriche svolte, sono stati ottenuti i risultati riportati nella tabella 3.3, relativi alla probabilità di osservare valori uguali o maggiori all'indice di posizione considerato. Generalizziamo la numerosità campionaria in modo da rendere i calcoli più convenienti: indicheremo n come $n = 2^s B$ dove $s = 0, 1, 2, 3$ e $B = \{25, 30, 35\}$. Definiamo la moda come:

$$\text{mod}_{\,n} = \lfloor log_2 n \rfloor = s + \lfloor log_2 B \rfloor.$$

Tabella 3.3: Probabilità di osservare valori uguali o maggiori alla moda, considerando le probabilità definite nel corollario 1.

			$n = B$	$n = 2B$	$n = 2^2B$	$n = 2^3B$
	mod $_n$		4	5	6	7
B=25	P($S_n =$	mod $_n$)	0.2980	0.2768	0.2645	0.2574
	P($S_n >$	mod $_n$)	0.5496	0.5441	0.5423	0.5419
	mod $_n$		4	5	6	7
B=30	P($S_n =$	mod $_n$)	0.2743	0.2601	0.2511	0.2456
	P($S_n >$	mod $_n$)	0.6255	0.6160	0.6119	0.6100
	mod $_n$		5	6	7	8
B=35	P($S_n =$	mod $_n$)	0.2784	0.2627	0.2542	0.2495
	P($S_n >$	mod $_n$)	0.4102	0.4139	0.4167	0.4185

Guardando a queste quantità, possiamo confermare quanto affermato in precedenza: la probabilità di osservare valori pari alla moda, ossia quella che nella pagina precedente abbiamo nominato come probabilità massima, decresce all'aumentare della numerosità. Ma i risultati sui quali è giusto porre l'attenzione sono quelli riguardanti la probabilità di osservare valori maggiori della moda. Si noti in particolare che questa quantità varia molto poco al variare di s, ma dipende piuttosto da B. Qualunque sia questo valore, però, possiamo osservare degli alti valori di
P($S_n >$ mod $_n$), che ci portano a concludere, dunque, come sia frequente osservare periodi di concordanza/discordanza lunghi in confronto alla ristretta numerosità

campionaria. Questo rispecchia quanto detto precedentemente a commento della tabella 3.1.

3.2 Un confronto con i risultati approssimati

Nel capitolo precedente, nella sezione dedicata 2.2, sono stati presentati i risultati approssimati, ma non la loro effettiva compatibilità con le distribuzioni reali. In questa sezione, dunque, metteremo a confronto le analisi numeriche ottenute nelle pagine precedenti, con le quantità calcolate sulla base delle formule approssimate. In particolare, poniamo anche qui l'attenzione sulla quantità S_n piuttosto che su Z_n, in quanto racchiude sia casi di concordanza che di discordanza.

Nell'equazione 2.8, abbiamo definito

$$\mathrm{P}(S_n = \lfloor \log_2(2^s B) \rfloor) \approx P_n(\lfloor \log_2(2^s B) \rfloor)$$

Dato che $\{\log_2 B\} \in [0, 1)$, segue direttamente che $P_n(\lfloor \log_2(2^s B) \rfloor)$ vari solo tra 0.233 e 0.250 . Considerando, quindi, i diversi valori di n presi sotto esame e sapendo che $P()$ è costante per ogni s, otteniamo:

$$\begin{aligned} P_n(\lfloor \log_2(2^s 25) \rfloor) &= 0.2482 \\ P_n(\lfloor \log_2(2^s 30) \rfloor) &= 0.2383 \\ P_n(\lfloor \log_2(2^s 35) \rfloor) &= 0.2438 \end{aligned}$$

Confrontando questi risultati con la tabella 3.3, possiamo osservare che non si discostano molto dalle cifre ottenute precedentemente e rispettano anche i rapporti di proporzione tra i diversi valori di B, infatti per $B = 25$ abbiamo ancora la probabilità più alta. Riportiamo poi l'approssimazione per la distribuzione di ripartizione:

$$\mathrm{P}(S_n > \lfloor \log_2(2^s B) \rfloor) \approx 1 - \exp\left(-2^{\{\log_2 B\}-1}\right)$$

e calcoliamo le quantità:

$$\mathrm{P}(S_n > \lfloor \log_2(2^s B) \rfloor) \approx \begin{cases} 0.5422 & \text{per } B = 25 \\ 0.6084 & \text{per } B = 30 \\ 0.4212 & \text{per } B = 35 \end{cases}$$

Anche in questo caso, abbiamo trovato valori simili a quelli calcolati con le formule esatte.

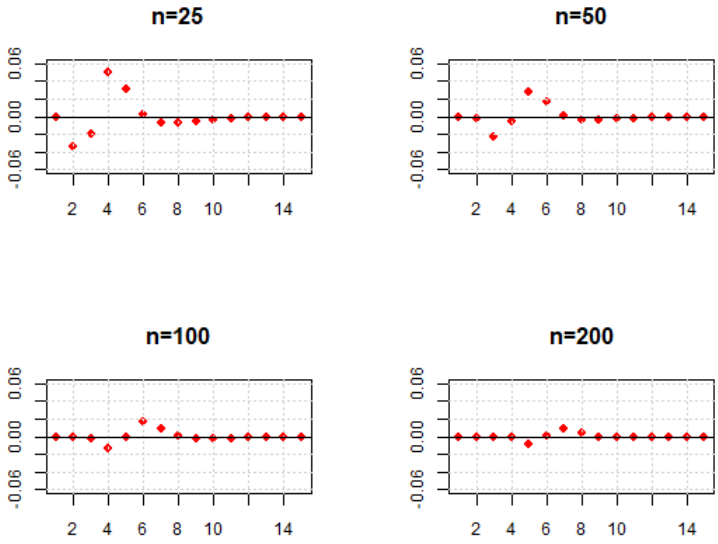

Figura 3.2: $\mathrm{P}(S_n = k) - P_n(k)\,, k = 1, \ldots, 15$

Nei grafici sopra riportati 3.2, estratti dallo studio numerico svolto, possiamo osservare in modo più evidente le differenze tra la probabilità esatta e quella approssimata, $\mathrm{P}(S_n = k) - P_n(k)$. Nei due grafici più in basso, si può notare che la curva dei punti oscilla di poco attorno allo zero, possiamo quindi affermare che per $n \geq 100$ è possibile ottenere delle buone approssimazioni. Per $n = 25$ e $n = 50$, invece, si osservano deviazioni dalle probabilità definite nel corollario 1 considerevoli.

Si confrontino, infine, i risultati approssimati ottenuti per μ_n con i valori attesi in 3.2. Vediamo che le quantità coincidono fino alla seconda cifra decimale. Possiamo concludere, quindi, che μ_n rappresenti una buona approssimazione del valore atteso di S_n.

Tabella 3.4: Confronto tra il valore atteso di S_n e la sua approssimazione μ_n.

	$n = 25$	$n = 50$	$n = 100$	$n = 200$
μ_n	4.9766	5.9766	6.9766	7.9766
$\mathrm{E}(S_n)$	4.9799	5.9783	6.9774	7.9770

Per quanto riguarda la varianza invece, abbiamo visto nel capitolo precedente che può essere approssimata da:

$$\mathrm{Var}(S_n) \approx \frac{\pi^2}{6\ln^2 2} + \frac{1}{12} \approx 3.5070\,.$$

Questo valore non dipende però, da n e non rispecchia neanche i valori reali della varianza riportati nella tabella 3.2. Non è possibile, quindi, considerarla una buona approssimazione.

Capitolo 4

Random walks correlati

Nei capitoli precedenti abbiamo preso in considerazione il caso di *random walks* indipendenti tra loro. In questo capitolo, vogliamo invece cambiare le ipotesi di partenza e capire a che risultati si giunge trattando il caso di *random walks* correlati.

4.1 Metodologia

In presenza di correlazione tra le due serie storiche, è necessario scartare l'ipotesi contenuta nell'assunzione 1 e considerare quindi $\mathrm{P}(C_i = 0) \neq \mathrm{P}(C_i = 1)$, al fine di ottenere un processo di Bernoulli asimmetrico (i.e., lancio di una moneta truccata). Diremo che maggiore è la correlazione tra i due *random walks*, maggiore è la probabilità di concordanza p:

$$p := \mathrm{P}(C_i = 0) \quad \text{e} \quad q := 1 - p = \mathrm{P}(C_i = 1).$$

Di conseguenza, una correlazione negativa implica $p < 1/2$.

Assunto 2. *Sia* $(\Delta X_i, \Delta Y_i)_{i=1,\dots,n}$ *una sequenza di variabili casuali continue, indipendenti e identicamente distribuite con* $0 < p < 1$.

Possiamo dedurne allora quanto segue:

Proposizione 3. *Sia* $(X_i, Y_i)_{i=1,\dots,n}$ *definito come nell'equazione 2.1, con le proprietà stabilite nell'assunto precedente 2. Per qualunque intero* z*, è possibile verificare che*

$$P(Z_n - \lfloor m_{n,p} \rfloor < z) = \exp\left(-p^{z-\lfloor m_{n,p} \rfloor}\right) + o(1)$$

dove $m_{n,p} := log_{1/p}(nq)$.

Dimostrazione. (Vedi teorema1, Gordon, Schilling e Waterman 1986). □

Con il risultato appena ottenuto, possiamo definire l'approssimazione per la funzione di ripartizione:

$$\mathrm{P}(Z_n \leq k) \approx \exp\left(-p^{\,k+1-m_{n,p}}\right). \tag{4.1}$$

Si noti che $m_{n,1/2} = log_2(n) - 1$ e $\{log_2(n) - 1\} = \{log_2 n\}$. La proposizione 2 risulta, quindi un caso particolare ($p = 1/2$) della formula più generale appena enunciata.

4.2 Studio di simulazione

Quando si tratta di *random walks* correlati, non abbiamo a disposizione una formula esatta per la funzione di ripartizione della *zero run*, quindi possiamo solo affidarci all'approssimazione stimata in 4.1. Per valutare dunque, l'attendibilità della formula calcolata, eseguiamo un piccolo studio di Monte Carlo su una base di 10^5 replicazioni.

Tabella 4.1: Confronto tra le probabilità $\mathrm{P}(Z_n > \lfloor \log_2 n \rfloor)$, costruite sull'approssimazione 4.1, e le stime di Monte Carlo ottenute su 10^5 simulazioni, per p variabile.

	n	$n = 25$	$n = 50$	$n = 100$	$n = 200$
	$\lfloor \log_2 n \rfloor$	4	5	6	7
Approssimazioni	$p = 0.23$	0.0123	0.0057	0.0026	0.0012
	$p = 0.42$	0.1726	0.1472	0.1252	0.1062
	$p = 0.50$	0.3234	0.3234	0.3234	0.3234
	$p = 0.58$	0.4980	0.5504	0.6044	0.6590
	$p = 0.77$	0.7891	0.0.9090	0.9751	0.9966
Stime di Monte Carlo	$p = 0.23$	0.0105	0.0052	0.0024	0.0010
	$p = 0.42$	0.1569	0.1355	0.1223	0.1025
	$p = 0.50$	0.3109	0.3139	0.3184	0.3191
	$p = 0.58$	0.5167	0.5634	0.6155	0.6670
	$p = 0.77$	0.9303	0.9770	0.9943	0.9996

Con riferimento alla tabella 4.1, possiamo commentare la bontà delle approssimazioni definite dalla formula 4.1. All'interno dello schema troviamo le probabilità di osservare una *zero run* di lunghezza superiore alla moda. Nella prima fascia le probabilità sono state calcolate facendo riferimento all'approssimazione,

$\mathrm{P}(Z_n > \lfloor \log_2 n \rfloor) = 1 - \exp\left(-p^{\lfloor \log_2 n \rfloor + 1 - m_{n,p}}\right)$, mentre le stime di Monte Carlo sono state ottenute sulla base di 10^5 simulazioni, generate a partire da $(\Delta X_i, \Delta Y_i)$ con distribuzione normale bivariata, con valori di partenza nulli e correlazione ρ variabile, $\rho \in \{-0.75, -0.25, 0, 0.25, 0.75\}$. Da questi indici di correlazione sono stati tratti i valori per p, $p = \{0.23, 0.42, 0.50, 0.58, 0.77\}$. Come si intuisce facilmente, per $p > 0.50$, ossia in presenza di correlazione positiva, aumenta la probabilità di identificare *zero run* più lunghe della moda e si può osservare come con $p = 0.77$ sia pari all'incirca di 0.93. Dall'altro lato, una $p < 0.5$ sta ad indicare che i due *random walks* tendono a variare in maniere discordante e di conseguenza la *zero run* diminuisce. Tuttavia, questo non sta a significare che la correlazione tra le due serie storiche diminuisca. Definiamo O_n come la lunghezza della più lunga sequenza di uno in $(C_i)_{i=1,\dots,n}$. Dall'approssimazione 4.1, si ottiene quindi

$$\mathrm{P}(O_n \leq k) \approx \exp\left(-q^{\,k+1-\mu_{n,q}}\right),$$

dove $\mu_{n,q}$ è definito analogamente alla quantità $m_{n,p}$: $\mu_{n,q} := \log_{1/q}(np)$.

La tabella 4.1 porta quindi ad affermare che maggiore è la correlazione tra i due *random walks*, o maggiore $|p - 0.5|$, più grandi saranno le *zero run* o le *uno run* in $(C_i)_{i=1,\dots,n}$, a seconda del segno della correlazione. Per quanto riguarda, invece, un confronto tra approssimazioni e stime di Monte Carlo, possiamo concludere che ci sia un buon adattamento per valori di $p \leq 0.50$, mentre per $p = 0.58$ e $p = 0.77$ i risultati approssimati tendono a essere più contenuti rispetto a quelli stimati con le simulazioni.

Conclusioni

In questa tesi viene trattato l'argomento della correlazione spuria tra due *random walks* indipendenti e non, concentrandosi su campioni di numerosità finita piccola, $n = 25, 50, 100, 200$. Per verificare che associazioni spurie rappresentino la regola piuttosto che l'eccezione, abbiamo calcolato la massima lunghezza di istanti consecutivi di concordanza o discordanza e abbiamo ottenuto, poi, la probabilità di osservare valori maggiori o uguali alla moda della lunghezza stessa.

Nella prima parte dell'elaborato è stata presentata una panoramica generale sulla storia della regressione spuria, fino ad arrivare agli studi condotti ai giorni nostri. A differenza di quanto fatto dagli Autori citati, in questa tesi non si parla di alcun tipo di stimatore o di erronea formulazione del modello di regressione, bensì si analizzano due *random walks* a partire da periodi di concordanza o discordanza tra essi.
Si è parlato, quindi, di *zero run* e *uno run*, rispettivamente come lunghezza della massima sequenza di istanti consecutivi in cui le due serie si muovono nella stessa direzione (concordanza), e lunghezza della massima sequenza di periodi consecutivi discordanti. Per queste quantità, sono state presentate le funzioni di densità e distribuzione a partire da sequenze di Fibonacci di ordine l, $(f_m^{(l)})_{m=1,2,\dots}$.
Con riferimento alle tabelle 3.2 e 3.3, possiamo dire che per $n = 25$, il valore atteso della massima lunghezza di un'associazione casuale (concordanza o discordanza consecutiva) è pari a 4.9799 e che la probabilità di osservare una sequenza maggiore di 4 è del 54.96%. Per $n = 30$, quest'ultimo valore si alza al 62.55%.

Per tutte le quantità descritte sopra, abbiamo calcolato le rispettive approssimazioni: per $B = 25$ e $B = 30$, la probabilità di osservare un valore maggiore alla moda è rispettivamente pari al 54.22% e al 60.84%. A proposito della funzione di densità, invece, una buona sintesi di valutazione delle approssimazioni si può osservare nei grafici 3.2. Notiamo che per numerosità campionaria piccola, la differenza tra valori teorici e approssimati causa una oscillazione maggiore della curva. Questa distorsione viene meno, invece, al crescere di n. Anche le approssimazioni del valore

atteso di S_n rappresentano bene i valori teorici: per $n = 25$, $\mathrm{E}(S_n) = 4.9799$ viene approssimato da $\mu_n = 4.9766$. Lo stesso non può essere detto, però, per il calcolo della varianza. Nonostante questa quantità, possiamo affermare di aver ottenuto delle buone approssimazioni dei risultati teorici.
Nell'ultimo capitolo viene presentata una piccola estensione al problema, rilasciando l'ipotesi di indipendenza e trattando serie correlate tra loro, con
$\rho = \{-0.75, -0.25, 0, 0.25, 0.75\}$. Non avendo a disposizione la distribuzione esatta, i risultati asintotici sono stati confrontati con una simulazione di Monte Carlo e si è giunti alla conclusione che maggiore è la correlazione tra *random walks*, maggiore è la probabilità di osservare *zero run* se il segno della correlazione è positivo, *uno run* se, invece, è presente una correlazione negativa. A conferma di questo, con riferimento alla tabella 4.1, possiamo dire che per $\rho = 0.75$ e $n = 25$, la probabilità di osservare sequenze di zeri maggiori della moda è pari a 78.91% se consideriamo l'approssimazione 4.1, 92.97% nelle stime di Monte Carlo. Considerando $n = 50$, otteniamo invece rispettivamente una probabilità del 90.90% e del 97.66%.

Rispetto alla numerosità campionaria, si osservano quindi dei lunghi periodi di concordanza o discordanza tra *random walks*, sia in caso di indipendenza che in presenza di correlazione e questo ci porta a giustificare la presenza di una correlazione campionaria priva di senso piuttosto alta tra passeggiate aleatorie.

Elenco delle tabelle

Elenco delle figure

Ringraziamenti

Questo spazio è dedicato a tutte le persone che mi sono state accanto durante la stesura di questa tesi e più in generale, durante l'intero percorso di studi.
In primis, grazie Giovanni per aver capito il mio modo di fare e la necessità di chiudermi in me stessa nei momenti di studio intenso. Grazie per continuare ad ascoltare le mie lamentele e minacciarmi con testate appena mi scoraggio un attimo. Ricordati che tutto il tempo che tolgo a te per dedicarmi allo studio, viene recuperato poi nella lista "Cose da fare finita la sessione".
Un grande grazie alla mamma per essermi sempre accanto, per prepararmi gli spuntini dolci per far passare la fame da nervoso e per provare a suggerire soluzioni ai problemi, anche se si sa che è solo un tirare ad indovinare perchè in realtà ti sembra che io studi geroglifici (motivo per cui è servito prendere appunti mentre spiegavo l'argomento di questa tesi). Grazie papà per allenarmi involontariamente a tenere a mente il calendario, chiedendomi sempre «Quando hai il prossimo esame?». Grazie Mattia per avermi abituato da sempre a studiare con la televisione accesa col volume al massimo: ha permesso di sviluppare il mio ascolto passivo. Grazie ai miei due fratelli a quattro zampe, Joker e Aki, per essere stati i compagni di passeggiata ideali per staccare dopo una giornata piena di studio. Gli altri studenti generalmente vanno in aula studio perchè a casa non riescono a studiare. Per me è esattamente il contrario: non riesco a studiare se sono fuori casa e questo è anche grazie all'ambiente che avete creato e create voi.
Un grazie, infine, alla professoressa Bisaglia per aver creduto in me e avermi dato la possibilità di continuare questa tesi, nonostante un viaggio di 20 giorni oltreoceano e un esame ancora da passare al limite con le scadenze.

Bibliografia

Aldrich, J. (1995). «Correlations Genuine and Spurious in Pearson and Yule». In: *Statistical Science* 10, pp. 364–376.

Box, G.E.P e G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

Engle, R.F. e C.W.J. Granger (1987). «Co-integration and error correction: representation, estimation and testing.» In: *Econometrica* 55, pp. 251–276.

Ernst, P.A., L.A. Shepp e A.J. Wyner (2017). «Yule's "Nonsense correlation" solved!» In: *The Annals of Statistics* 45, pp. 1789–1809.

Escudero, W.S. (2000). «A Primer on Unit-Roots and Cointegration». Tesi. Universidad Nacional de La Plata.

Földes, A. (1975a). «On the limit distribution of the longest heads.» In: *Matematikai Lapok* 26, pp. 105–116.

— (1975b). «The limit distribution of the length of the longest head-run». In: *Period Math Hung* 10, pp. 301–310.

Galton, F. (1888). «Co-relations and their measurement». In: *Proc. Roy. Soc. London Ser.* 41, pp. 135–145.

Gordon, L., M.F. Schilling e M.S. Waterman (1986). «An extreme value theory of long head runs.» In: *Probab Theory Relat Fields* 72, pp. 279–287.

Granger, C.W.J. (1981). «Some properties of time series data and their use in econometric model specification.» In: *Journal of Econometrics* 16, pp. 121–130.

Granger, C.W.J. e P. Newbold (1974). «Spurious regressions in econometrics». In: *J.Econometrics* 2, pp. 111–120.

Guibas, L.J. e A.M. Odlyzko (1980). «Long repetitive patterns in random sequences.» In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete.* 53, pp. 241–262.

Hassler, U. e M. Hosseinkouchack (2022). «Understanding nonsense correlation between (indipendent) random walks in finite sample». In: *Statistical Papers* 63, pp. 181–195.

Hendry, D.F., J.E.H. Davidson, F. Srba et al. (1977). «Econometric modeling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom». In: *Economic Newspaper* 88, pp. 661–692.

Hendry, D.F. e K. Juselius (2000). «Explaining Cointegration Analysis: Part 1». In: *The Energy Journal* 21, pp. 1–42.

Hooker, R. H. (1901). «On the correlation of the marriage-rate with trade». In: *J. Roy. Statist. Soc. Ser.* 64, pp. 485–492.

— (1905). «On the correlation of successive observa- tions illustrated by corn prices.» In: *J. Roy. Statist. Soc. Ser.* 68, pp. 696–703.

Johansen, S. (1988). «Statistical analysis of cointegration vectors.» In: *Journal of Economic Dynamics and Control* 12, pp. 231–254.

Klein, J.L. (1997). *Statistical Visions in Time: A History of Time Series Analysis, 1662-1938*. Cambrige, U.K.: Cambridge University Press.

Meuriot, V. (2015). «The concept of cointegration: the decisive meeting between Hendry e Granger (1975).» In: *Cahiers d'économie Politique* 68, pp. 91–118.

Pearson, K. (1892). *The Grammar of Science*. London: Walter Scott.

— (1896). «Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia». In: *Trans. Roy. Soc. London Ser.* 187, pp. 253–318.

— (1897). «On a form of spurious correlation which may arise when indices are used in the measurement of organs.» In: *Proc. Roy. Soc. London Ser.* 60, pp. 489–498.

Pearson, K., A. Lee e L. Bramley-Moore (1899). «Genetic (reproductive) selection: inheritance of fertility in man, and of fecundity in thoroughbred racehorses». In: *Philos. Trans. Roy. Soc. London Ser.* 192, pp. 257–330.

Pearson, K., A. Lee e E.M. Elderton (1910). «On the correlation of death rates.» In: *J. Roy. Statist. Soc. Ser.* 73, pp. 534–539.

Phillips, P.C.B. (1986). «Understanding spurious regressions in econometrics». In: *Journal of Econometrics* 33, pp. 311–340.

Révész, P. (1990). «Regularities and irregularities in a random 0, 1 sequence.» In: *Stat Pap* 31, pp. 95–101.

Sargan, J.D. (1964). «Wages and prices in the United Kingdom: A study in econometric methodology (with discussion)». In: *Econometric Analysis for National Economic Planning* 16, pp. 25–63.

Simon, H.A. (1954). «Spurious correlation: a causal interpretation». In: *J. Amer.Statist. Assoc.* 49, pp. 467–492.

Simpson, E.H. (1951). «The interpretation of interaction in contingency tables». In: *J. Roy. Statist. Soc. Ser.* 13, pp. 238–241.

Spickerman, W.R. e R.N. Joyner (1984). «Binet's formula for the recursive sequence of order K.» In: *Fibonacci Q* 22, pp. 327–331.

Student (1908). «Probable error of a correlation coefficient.» In: *Biometrika* 6, pp. 302–310.

Yule, G.U. (1903). «Notes on the theory of association of attributes in statistics.» In: *Biometrika* 2, pp. 121–134.

— (1910). «On the interpretation of correlations between indices or ratios.» In: *J. Roy. Statist. Soc. Ser.* 73, pp. 644–647.

— (1921). «On the time-correlation problem, with special reference to the variate-difference correlation method.» In: *J. Roy. Statist. Soc. Ser.* 84, pp. 497–526.

— (1926). «Why do we sometimes get nonsense-correlations between time series? A study in sampling and the nature of time-series». In: *J. Roy. Statist. Soc* 89, pp. 1–63.

— (1927). «On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers». In: *Philos. Trans. Roy. Soc. London Ser.* 226, pp. 267–298.