UNIVERSITA' DEGLI STUDI DI PADOVA

Facoltà di Scienze Statistiche

Corso di Laurea Triennale in Statistica,
Economia e Finanza

TESI DI LAUREA

ANALISI STATISTICA SULLA LEGGE DI BENFORD

Relatore: Prof. Francesco Lisi

Laureando: Massimo Schiavo

ANNO ACCADEMICO 2009/2010

- Indice.

1. Introduzione.

1.1. Genesi della legge di Benford e sue spiegazioni in letteratura.

"In una sequenza casuale di numeri la probabilità che uno di essi inizi per 1 o per 9 è praticamente la stessa". Fino alla fine del diciannovesimo secolo questa affermazione aveva un senso e non aveva mai trovato alcuna confutazione. Tuttavia Newcomb (1881) fece un'osservazione casuale, che il progresso tecnologico non avrebbe più permesso. Bisogna tenere a mente che quando i computer non esistevano i calcoli si facevano a mano. Per semplificare almeno un po' la vita di chi aveva a che fare tutti i giorni con centinaia e centinaia di calcoli venivano utilizzati i logaritmi, che permettono di trasformare le moltiplicazioni in addizioni, al prezzo di consultare le "tavole dei logaritmi", che danno la conversione tra un numero ed il suo logaritmo. Newcomb, che come tutti gli astronomi aveva bisogno di fare moltissimi calcoli, si accorse che i bordi delle prime pagine di un manuale contenente delle tabelle logaritmiche erano più sporchi (e quindi più utilizzati) di quelli delle ultime pagine. Sembrava che gli capitasse più spesso di cercare il logaritmo di un numero che iniziava con una cifra "piccola". Newcomb, dopo aver annotato ciò, non se ne occupò più.

Una cinquantina di anni dopo il fisico Frank Benford, che all'epoca lavorava all'interno dei Research Laboratories della General Electric, decise di migliorare con delle evidenze empiriche quanto aveva casualmente notato Newcomb anni prima.

Benford (1938) raccolse all'incirca ventimila dati relativi a ventuno elementi tra di loro assai diversi, misurando la frequenza delle cifre da 1 a 9 compresi, non tenendo in considerazione lo 0. Quello che riuscì ad ottenere fu che le frequenze delle nove cifre prese in considerazione non erano equamente ripartite. Infatti, ad esempio, notò che mentre la probabilità che la prima cifra fosse pari a 1 era mediamente del 30,6%, quella che la prima cifra fosse pari a 9 era mediamente pari al 4,7%.

Benford non riuscì a spiegare questo fenomeno, ma formulò una legge, che da quel momento prese il nome di legge di Benford o legge della prima cifra. Essa può essere così espressa:

$$\Pr(\text{prima cifra} = d) = \log_{10}(d + 1) - \log_{10}(d) = \log_{10}(1 + 1 / d) \quad \text{dove d} = 1,..,9.$$

Da questa formula si ottengono le frequenze espresse in Tabella 1.

Prima cifra significativa	Frequenza relativa
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

Tabella 1. Frequenze relative della prima cifra significativa nella legge di Benford.

Della legge di Benford esiste anche una variante che si applica alla seconda cifra significativa:

$$\sum_{k=1}^{9} \log_{10}\left(1+\left(\frac{1}{10k+d}\right)\right) \quad \text{dove } d = 0,1,...,9.$$

Una spiegazione dei risultati ottenuti da Benford fu data da Pinkham (1961) da un punto di vista puramente logico. Egli provò ad applicare la legge di Benford ad alcuni elementi naturali, tra cui la lunghezza dei fiumi americani. Sulla base dei risultati ottenuti affermò che, se esiste una legge che disciplina la prima cifra dei numeri relativi, ad esempio, ad alcuni fenomeni naturali, deve valere indipendentemente dalle unità di misura adottate, deve essere cioè invariante di scala. Osservò infatti che se, ad esempio, si misura la lunghezza dei fiumi americani in chilometri o in miglia cambiavano i singoli numeri, ma le frequenze relative della prima cifra significativa erano le stesse. Pinkham ha quindi dimostrato l'invarianza di scala della legge di Benford e, in particolare, che essa è l'unica legge sulla distribuzione delle prime cifre che gode di questa proprietà.

Trent'anni dopo che Pinkham aveva interprato i risultati ottenuti da Benford, Hill (1995) diede la prima spiegazione rigorosa della legge di Benford. Egli elaborò la condizione sufficiente affinché dei dati casuali seguissero detta legge: se si scelgono in modo casuale delle distribuzioni, che godono della proprietà di essere "scale or base invariant", e dei campioni casuali sono presi da ciascuna di queste distribuzioni, le frequenze relative della prima cifra significativa dei numeri che si ottengono unendo i campioni casuali seguiranno la legge di Benford.

Fewster (2009) ha osservato che molte ricerche condotte fino ad allora sulla legge di Benford evidenziavano che, anche se spesso la legge era rispettata, le spiegazioni a supporto di ciò non erano soddisfacenti. L'autore, anche tramite procedure grafiche, ha fornito una delucidazione semplice ed intuitiva del perché e del quando la legge è applicabile. Per verificare se dei dati casuali estratti da una qualsiasi distribuzione sono conformi alla legge di Benford (condizione che l'autore indica con "benfordness") bisogna controllare il supporto della distribuzione di $\log_{10}$: se tale supporto è sufficientemente ampio, allora ci si dovrebbe attendere "benfordness".

1.2. Applicazioni della legge di Benford.

Varian (1971) suggerì la possibilità di utilizzare la legge di Benford per individuare eventuali falsificazioni nelle raccolte di dati utilizzate per il supporto a decisioni politiche. Egli, basandosi sul presupposto che chi vuole "addomesticare" i dati ha una preferenza ad utilizzare numeri distribuiti in modo non "naturale" (che non seguono cioè la legge di Benford), propose di confrontare la frequenza relativa della prima cifra dei numeri utilizzati per il supporto a decisioni politiche con quelle "teoriche" della legge di Benford, in modo tale da evidenziare eventuali risultati anomali. Nel 2009, partendo da quanto aveva proposto Varian, il ministro degli interni iraniano Boudewijn F. Roukema, in seguito alla denuncia di brogli elettorali fatta da di Moussavi (leader dell'opposizione) contro il presidente uscente Ahmadinejad, ha deciso di controllare se vi fossero state delle irregolarità nel conteggio dei voti. Città per città egli ha analizzato le frequenze relative della prima

cifra significativa del numero di voti e ha confrontato le frequenze da lui ottenute con quelle "teoriche" della legge di Benford. Come risultato, dimostrante la veridicità delle accuse di Moussavi, Roukema ha notato che mentre tutte le altre cifre avevano delle frequenze molto simili a quelle "teoriche", il numero 7 compariva troppe volte. Deciso ad approfondire la questione, il ministro ha notato che l'anomalia riguardava tre delle sei aree più grandi dell'Iran. In queste zone il vincitore Ahmadinejad aveva ottenuto una percentuale di voti molto più elevata che nelle altre aree.

L'applicazione probabilmente più conosciuta della legge fu quella fatta da Nigrini (1996). Dopo aver applicato e testato l'efficacia della legge di Benford su casi reali di frode accertata nel 1992, egli se ne servì per testare la credibilità delle dichiarazioni dei redditi. Per fare ciò elaborò dei programmi che gli permisero di individuare distribuzioni numeriche sospette nelle dichiarazioni dei redditi. In Tabella 2 sono rappresentate le frequenze della prima cifra significativa ottenute da Nigrini nel 1996 analizzando le dichiarazioni dei redditi di alcuni Stati americani. Appare evidente il divario esistente tra le percentuali delle dichiarazioni "corrette" e quelle fraudolente.

	1	2	3	4	5	6	7	8	9
Legge di Benford	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.5%
Dichiarazioni "corrette"	30.5%	17.8%	12.6%	9.6%	7.8%	6.6%	5.6%	5.0%	4.5%
Dichiarazioni evasori	0%	1.9%	0%	9.7%	61.2%	23.3%	1.0%	2.9%	0%

Tabella 2. Nelle tre righe sono riportate le frequenze relative della prima cifra significativa, nella prima quelle "teoriche" della legge di Benford (riportate anche in Tabella 1), nella seconda quelle ottenute dalle dichiarazioni dei redditi "corrette" e nell'ultima quelle relative alle dichiarazioni dei redditi degli evasori.

Vista la sua efficacia, la procedura proposta da Nigrini è oggi utilizzata nella maggior parte degli Stati americani (tra cui la California) per contribuire ad individuare possibili evasori.

La legge di Benford può essere anche applicata ad uno o più indici azionari. In particolare Ley (1996) provò a verificare se la distribuzione delle frequenze relative della prima cifra significativa dei rendimenti giornalieri di due indici azionari americani (ovvero lo S&P per il periodo 1926-1993 e l'indice Dow Jones per il periodo 1900-1993), seguisse o meno la legge di Benford: come risultato ottenne che entrambi gli indici azionari rispettavano la suddetta legge. Ley, come spiegazione dell'aderenza dei rendimenti giornalieri alla legge di Benford, asserì che, nel caso dell'indice Dow Jones, ciò era dovuto al fatto che variazioni di piccola entità dei prezzi (che avvengono in un ristretto arco temporale) sono più probabili di quelle di entità superiore .

Infine, Corazza et al. (2010) hanno applicato la legge di Benford a 3067 rendimenti logaritmici giornalieri di 361 differenti titoli per il periodo che va dal 14 Agosto 1995 al 17 Ottobre 2007, tutti appartenenti all'indice S&P 500. Hanno ottenuto che per il periodo considerato, da un punto di vista grafico, i rendimenti rispettavano la suddetta legge, fatto che gli autori hanno interpretato come sinonimo di efficienza dell'indice preso in considerazione. Dalle loro analisi inoltre si è notato che i rendimenti non rispettavano la legge di Benford soprattutto in corrispondenza di eventi particolarmente negativi per il mercato azionario, e non solo, come l'11 Settembre 2001.

1.3. Svolgimento tesi.

L'elaborato finale si può suddividere in due parti.

Nella prima si verificherà se la distribuzione delle frequenze relative della prima cifra significativa di dati ottenuti tramite simulazione segua o meno la legge di Benford. I suddetti dati saranno creati in due differenti modalità. Inizialmente verranno estratte un certo numero di osservazioni da una distribuzione predefinita. In un secondo momento, per tentare di verificare la "condizione sufficiente" descritta da Hill, il campione casuale sarà formato estraendo un fissato numero di dati casuali da cinque differenti distribuzioni. Dopo aver verificato se i campioni casuali ottenuti in entrambi i modi rispettino o meno la legge di Benford, si verificherà se essi seguano una versione "approssimata" di suddetta legge.

Nella seconda parte si analizzerà se le frequenze relative della prima cifra significativa dei rendimenti logaritmici di tre differenti indici azionari seguano o meno la legge di Benford. Anche in questo caso in un secondo momento si verificherà cosa accade se si confrontano i rendimenti con una versione "approssimata" della legge di Benford.

2. Analisi di serie simulate.

Per verificare se la distribuzione delle frequenze relative della prima cifra significativa dei numeri ottenuti tramite simulazione segue o meno la legge di Benford è stata utilizzata la procedura di seguito descritta. Definito X il vettore di lunghezza pari ad n (numero di dati casuali del campione preso in considerazione), contenente le prime cifre significative dei numeri ottenuti tramite simulazione, ossia $X = (x_1, x_2, ..., x_n)$, la sua vera distribuzione si può indicare con $F_X(x)$, quindi $X \sim F_X(x)$. Definita poi F^B la distribuzione sottostante alla legge di Benford, per poter verificare in modo empirico se i dati simulati rispettano o meno la suddetta legge, è stato utilizzato il seguente sistema di ipotesi:

H0: $F_X(x) = F^B$

H1: $F_X(x) \neq F^B$

Per verificarlo è stato impiegato il seguente test per la bontà di adattamento del tipo test chi-quadrato di Pearson:

$$\chi^2 = \sum_{i=1}^{9} \frac{(f_i^{oss} - f_i^B)^2}{f_i^B} \qquad (1)$$

dove f_i^{oss} sono le frequenze della prima cifra significativa misurate sui dati simulati ed invece f_i^B sono le frequenze "teoriche" della legge di Benford, riportate in Tabella 1.

E' da notare che per numerosità campionarie elevate e sotto l'ipotesi nulla H0 il test (1) ha una distribuzione asintotica chi-quadrato con otto gradi di libertà, ossia:

$\chi^2 \sim \chi_8^2$ per numerosità campionarie elevate e sotto H0.

Dopo aver verificato se i dati ottenuti tramite simulazione rispettano o meno la legge di Benford, si è provato a verificare cosa accade se si impiega una versione "approssimata" della suddetta legge. Per poterla ottenere si è proceduto nel modo seguente:

definita P_i^B (i=1,...,9) la probabilità di ottenere la i-esima cifra come prima cifra significativa di un qualsiasi numero casuale, si considera la probabilità "perturbata", data da:

$$P_i^{B^\alpha} = P_i^B + \varepsilon_i$$

dove $\varepsilon_i \sim N(0, (\alpha * P_i^B)^2)$, ε_i *indipendenti*.

Si sono costruite delle probabilità "perturbate", che però in media non sono altro che: $E[P_i^{B^\alpha}] = P_i^B$

Dopo aver definito la frequenza (dove n è la numerosità del campione casuale analizzato):

$$f_i^{B^\alpha} = P_i^{B^\alpha} * n,$$

si nota che le frequenze ottenute in questo modo dipendono dalle probabilità "perturbate", a loro volta dipendenti dal termine di errore ε_i ; si pone il problema di come formare le suddette frequenze. Una possibile soluzione è quella di ricorrere a metodi di simulazione, cioè per j = 1, ..., m sono state ripetute le seguenti operazioni:

a. seguendo la loro distribuzione si estraggono gli ε_i, da cui si ottengono le seguenti probabilità:

$$P_i^{B^{\alpha^*}} \quad \forall \, i;$$

b. si normalizzano le probabilità ottenute affinché sommino ad 1:

$$P_i^{B^\alpha} = \frac{P_i^{B^{\alpha^*}}}{\sum_{i=1}^{9} P_i^{B^{\alpha^*}}};$$

c. si costruiscono le frequenze nel modo seguente (n è sempre la numerosità campionaria):

$$f_i^{B^\alpha} = P_i^{B^\alpha} * n;$$

d. si inseriscono le frequenze ottenute nel seguente test "perturbato", che nella forma è simile al test (1), ma non nella sostanza:

$$\chi_j^2 = \sum_{i=1}^{9} \frac{\left(f_i^{B^\alpha} - f_i^B\right)^2}{f_i^B} \qquad (2)$$

f_i^B sono anche in questo caso le frequenze "teoriche" della legge di Benford.

Dopo m ripetizioni di tale procedura si hanno m valori di χ_j^2, j = 1, ..., m, a partire dai quali ci si può calcolare il valore critico della statistica (2), indicato con $\chi_\gamma^{2\,simulato}$.
A questo punto si può utilizzare ancora la formula (1), riferendosi tuttavia al valore critico $\chi_\gamma^{2\,simulato}$.

2.1. Estrazione di dati casuali da una distribuzione.

Inizialmente i dati casuali sono stati generati da una singola distribuzione: in un primo momento dalla gaussiana di media nulla e varianza unitaria (gaussiana standardizzata), in seguito dalla uniforme (0,1) ed infine dalla distribuzione esponenziale di parametro λ pari a 2. Si è poi verificato se i dati generati rispettassero o meno la legge di Benford con l'ausilio della statistica (1), ripetendo tale procedura per mille volte e verificando cosa accadeva al variare del numero di dati estratti da una singola distribuzione, partendo prima con 100 dati estratti, per passare poi a 250, 500, 1000 ed infine 1500. Dopo aver effettuato ciò si è verificato cosa accadeva se si utilizzava una versione "approssimata" della legge di Benford, ottenuta con la procedura descritta nel paragrafo precedente. Anche in questo caso si è verificato cosa si otteneva al variare della numerosità campionaria (prima pari a 250, poi a 500 ed infine a 1000) e del termine di errore α (prima pari a 0.025, poi a 0.05 ed infine a 0.1), utilizzato nella distribuzione degli errori ε_i.

Nel Grafico 1 sono riportate le percentuali di accettazione ottenute al variare della numerosità campionaria n per le tre distribuzioni utilizzate nella generazione dei dati:

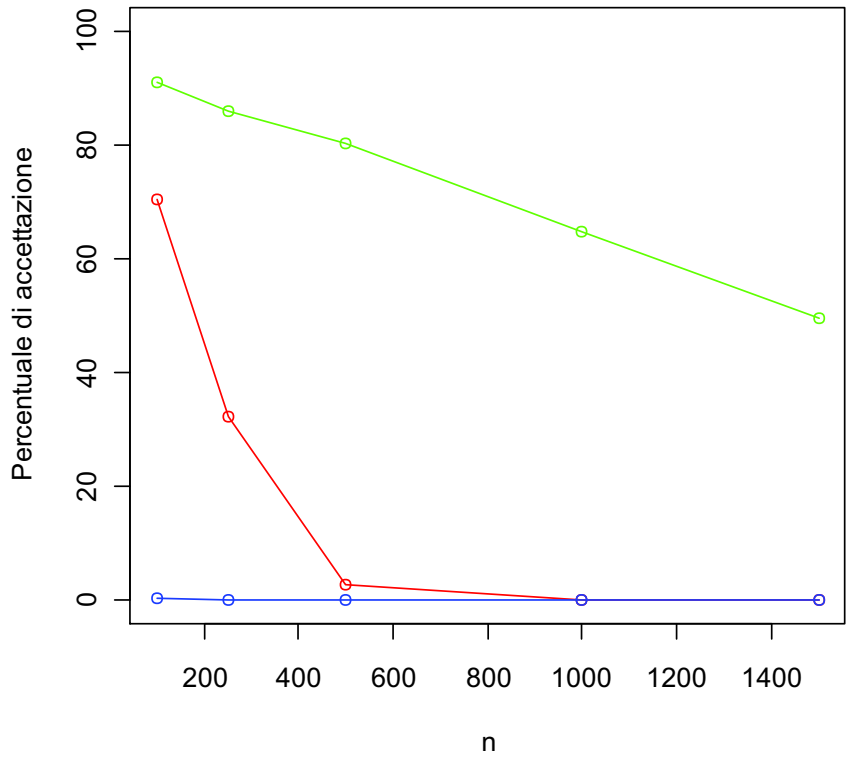

Grafico1. Percentuali di accettazione del test (1) al variare della numerosità campionaria n. Le percentuali di accettazione ottenute generando i dati dalla distribuzione uniforme sono rappresentate con la linea blu, quelle ottenute generandoli dalla distribuzione normale dalla linea rossa ed infine la linea verde raffigura le percentuali ottenute con la distribuzione esponenziale.

Quello che si può notare dal Grafico 1 è innanzitutto che nella distribuzione uniforme le percentuali di accettazione sono prossime allo zero per qualsiasi numerosità campionaria (infatti solo con 100 dati si riesce ad accettare l'ipotesi nulla ed in soli tre casi su mille). Nel caso invece dei dati generati dalla distribuzione normale si ha una percentuale di accettazione che aumenta al diminuire della numerosità campionaria. Se questo fatto si può spiegare con una perdita di potenza del test (1), bisogna tuttavia notare come con 1500 e 1000 dati (quindi in corrispondenza di numerosità campionarie piuttosto elevate) la percentuale di accettazione sia praticamente pari a zero. Estraendo i dati dalla distribuzione esponenziale, le suddette percentuali sono sempre al di sopra del 50% per qualsiasi numerosità campionaria, arrivando a superare il 90% con 100 dati.

Nel caso in cui si estraggano dati casuali da una sola distribuzione quello che si ottiene è un sostanziale non rispetto della legge di Benford da parte dei dati generati dalla distribuzione normale e da quella uniforme, mentre nel caso di quelli ottenuti dalla distribuzione esponenziale si ottengono delle percentuali di accettazione che possono far affermare che le osservazioni estratte da tale distribuzione seguono la suddetta legge.

Dopo questa prima serie di operazioni, l'analisi si è concentrata su una versione della legge "approssimata". Nei Grafici 2, 3 e 4 sono riportate le percentuali di accettazione ottenute rispettivamente dalla distribuzione normale, uniforme ed esponenziale al variare della numerosità campionaria dei dati casuali (prima con 250 osservazioni, poi con 500 ed infine con 1000) e del termine α (prima pari a 0.025, poi a 0.05 ed infine a 0.1).

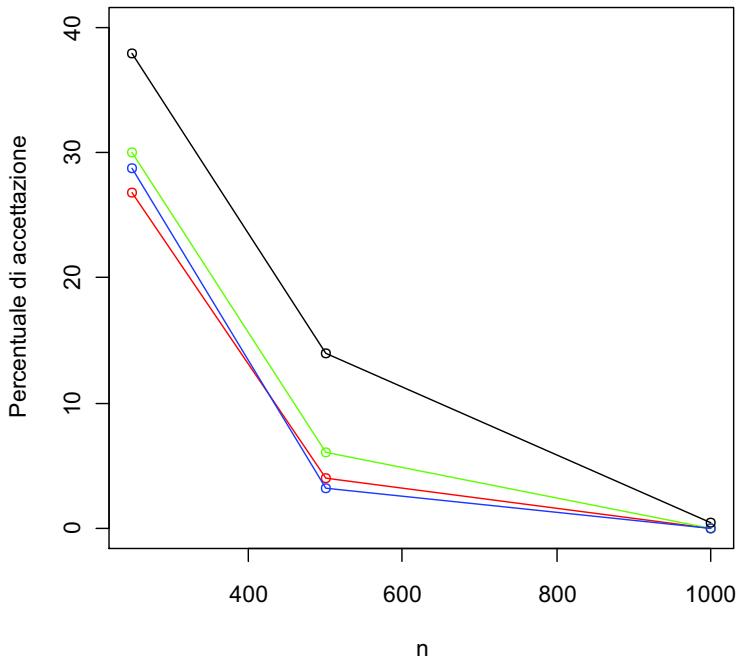

Grafico 2. Percentuali di accettazione del test (1) per dati originati dalla distribuzione normale al variare della numerosità campionaria n e del parametro α. La linea rossa rappresenta le percentuali di accettazione nel caso in cui il suddetto parametro valga zero (senza "perturbazione"), quella blu quando α vale 0.025, quella verde quando vale 0.05 ed infine quella nera quando vale 0.1.

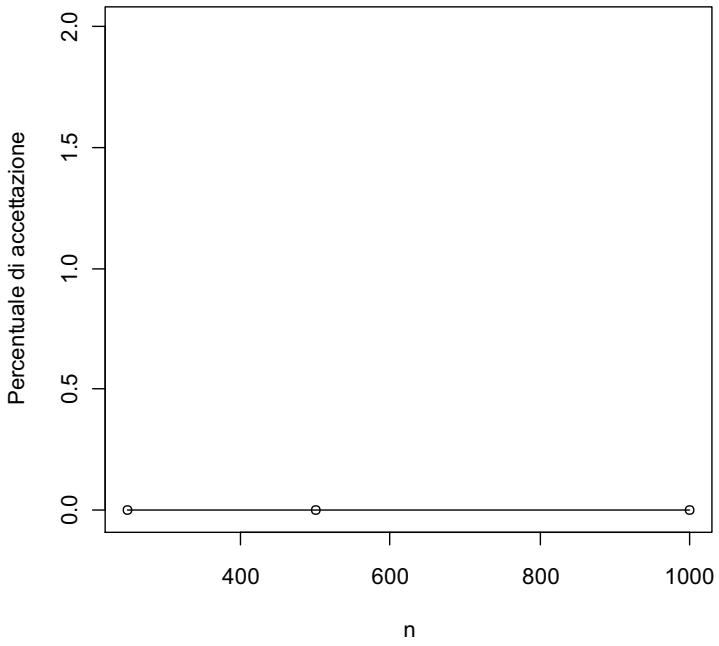

Grafico 3. Percentuali di accettazione del test (1) per dati originati dalla distribuzione uniforme al variare della numerosità campionaria n e del parametro α . In questo caso per qualsiasi valore del suddetto parametro (pari a zero, 0.025, 0.05, 0.1) le quattro linee coincidono.

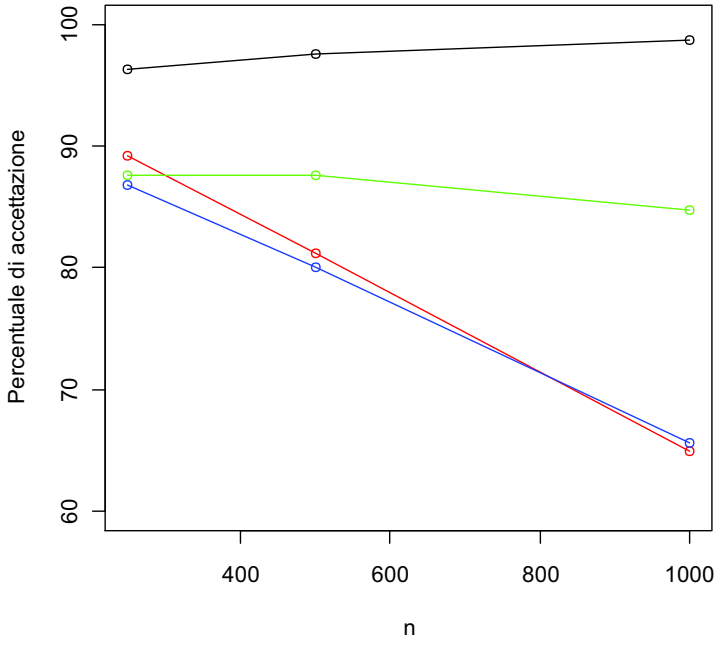

Grafico 4. Percentuali di accettazione del test (1) per dati originati dalla distribuzione esponenziale al variare della numerosità campionaria n e del parametro α . La linea rossa rappresenta le percentuali di accettazione nel caso in cui il suddetto parametro valga zero (senza "perturbazione"), quella blu quando α vale 0.025, quella verde quando vale 0.05 ed infine quella nera quando vale 0.1.

Come si può vedere dal Grafico 3 nel caso della distribuzione uniforme, in corrispondenza di qualsiasi numerosità campionaria e qualsiasi valore del parametro α, i dati generati non seguono la legge di Benford.

Nel Grafico 2 si vede come nel caso della distribuzione normale all'aumentare del valore di α ed al diminuire della numerosità n aumenti anche la percentuale di accettazione; se il parametro vale 0.025 e 0.05 si ottengono delle percentuali praticamente identiche a quelle che si hanno nel caso di "assenza di perturbazione". Se si assume un valore del suddetto parametro pari a 0.1 si accentua ancora di più la differenza delle percentuali di accettazione con quelle della distribuzione senza di esso. E' da notare che per qualsiasi valore di α la suddetta percentuale rimane sempre pari a zero quando il campione è costituito da mille dati.

Le differenze più evidenti si notano nel caso di dati generati dalla distribuzione esponenziale. Infatti già senza alcun grado di perturbazione si hanno delle percentuali di accettazione molto elevate. Mentre con il valore del parametro pari a 0.025 si ottengono quasi le stesse percentuali che si ottenevano senza perturbazione, già con α pari a 0.5 si può notare una notevole differenza rispetto alla distribuzione con α pari a zero, in particolare per le numerosità campionarie elevate. Infine per un valore del suddetto parametro pari a 0.1 per qualsiasi numerosità campionaria la suddetta percentuale è sempre superiore al 95%.

Concludendo si può dire che, mentre nel caso di dati generati dalla distribuzione normale e da quella uniforme introducendo qualsiasi grado di perturbazione non si ha un'elevata aderenza di questi ultimi alla legge di Benford, nel caso di osservazioni estratte dalla distribuzione esponenziale, aggiungendo diversi gradi di perturbazione, si ha un'ulteriore conferma dell'aderenza dei dati originati da tale distribuzione alla suddetta legge.

Quello che si è ottenuto estraendo delle osservazioni da una singola distribuzione è che di fatto i risultati variano, e di molto, al variare della distribuzione scelta. Probabilmente, soprattutto nel caso della distribuzione normale e di quella uniforme, se si fosse attribuito un valore differente ai parametri, sarebbe potuta aumentare anche la percentuale di accettazione.

2.2. Estrazione di dati casuali da più distribuzioni.

Dopo aver visto come estraendo dati casuali da una sola distribuzione si ottenevano dei risultati discordanti, si è provato a verificare cosa accade con un campione ottenuto generando un numero via via differente di dati da diverse distribuzioni casuali.

Le distribuzioni casuali da cui si è estratto un numero k via via differente di dati sono state le seguenti: la distribuzione normale standardizzata, la distribuzione chi-quadrato con dieci gradi di libertà (dividendo i risultati per dieci in modo tale da non ottenere valori troppo elevati), la distribuzione beta con parametro α pari a 0.6 e β a 0.6, la distribuzione esponenziale con parametro λ pari a 3 ed infine la distribuzione normale con media pari a zero e varianza pari a 0.4.

Per verificare la "condizione sufficiente" esposta da Hill nel proprio articolo è stato preso un numero k di dati casuali (k=1, 10, 50, 100) da una distribuzione tra le cinque sopraelencate, per poi estrarre un egual numero k da un'altra distribuzione, ed il tutto in modo continuo per andare a formare un campione casuale costituito da n dati casuali (n=250, 500, 1000); tale procedura è stata ripetuta per cento volte.

Nel Grafico 5 sono esposte le percentuali di accettazione al variare dei dati casuali k estratti da ciascuna distribuzione e della numerosità campionaria n del campione con essi formato.

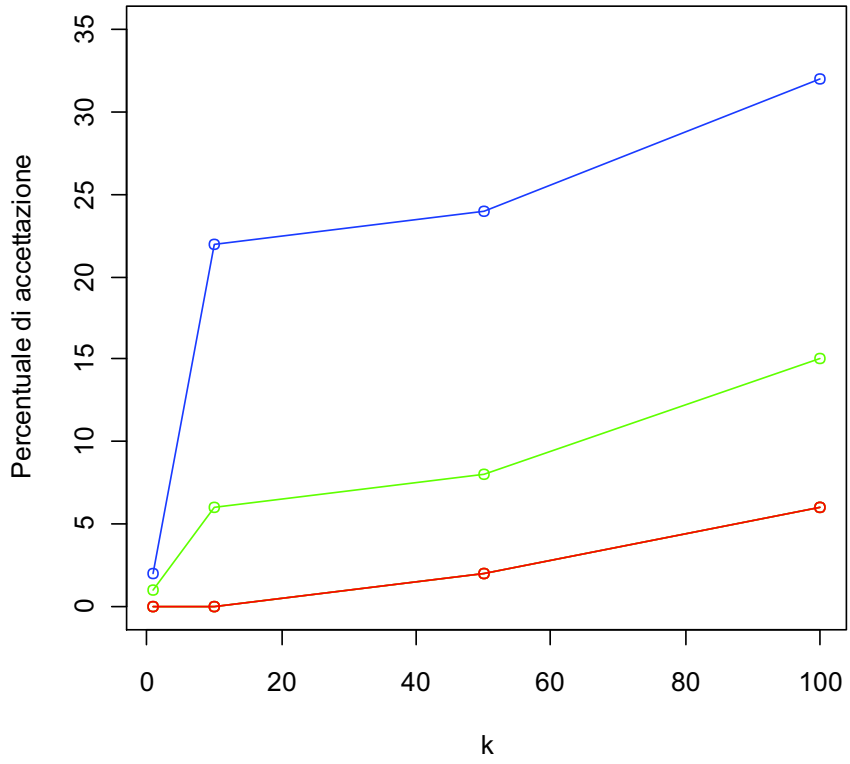

Grafico 5. Percentuali di accettazione del test (1) per dati originati da cinque differenti distribuzioni al variare della numerosità campionaria n e del numero di dati k estratti da ciascuna distribuzione. La linea rossa rappresenta le percentuali di accettazione per il campione casuale di numerosità n pari a 1000, la linea verde quello avente n pari a 500 e quella blu quello avente n pari a 250.

Dal Grafico 5 si può vedere che indipendentemente dalla numerosità campionaria n, se aumenta il numero di osservazioni k estratte da una singola distribuzione, aumenta anche la percentuale di accettazione del test (1). E' tuttavia da notare che la distribuzione delle frequenze relative della prima cifra significativa di campioni ottenuti estraendo un numero via via differente di dati da distribuzioni diverse non rispetta la legge di Benford, infatti la percentuale di accettazione più elevata (del 32%) si ottiene estraendo 100 dati casuali da una singola distribuzione per formare un campione di numerosità pari a 250.

Dai risultati ottenuti sembrerebbe che la condizione sufficiente affermata da Hill nel proprio articolo in questo caso non venga rispettata. Probabilmente ciò è dovuto al fatto che cinque distribuzioni casuali dalle quali estrarre i dati sono troppo poche, oppure che i valori attribuiti ai parametri delle singole distribuzioni non sono corretti. Infine, potrebbe essere che una o più tra le distribuzioni scelte non goda della proprietà di essere "scale or base invariant".

Anche in questo caso si è provato a verificare cosa accade se si considera una versione "approssimata" della legge di Benford.

Nei Grafici 6, 7, 8 sono riportate le percentuali di accettazione ottenute al variare della numerosità campionaria n e del numero di dati estratti dalla singola distribuzione k, con e senza il parametro di errore α, assunto pari a 0.05.

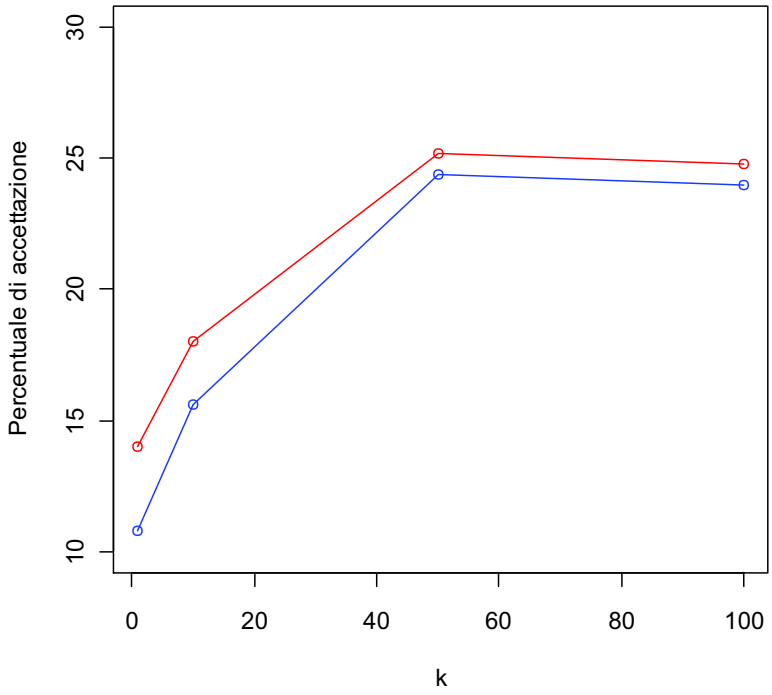

Grafico 6. Percentuali di accettazione del test (1) per un campione di numerosità campionaria pari a 250, al variare del numero k di dati estratti dalla singola distribuzione e del parametro α, assunto prima pari a zero e poi a 0.05. Mentre la linea blu rappresenta le percentuali di accettazione nel caso in cui il parametro di errore valga zero, quella rossa rappresenta le percentuali se il suddetto parametro vale 0.05.

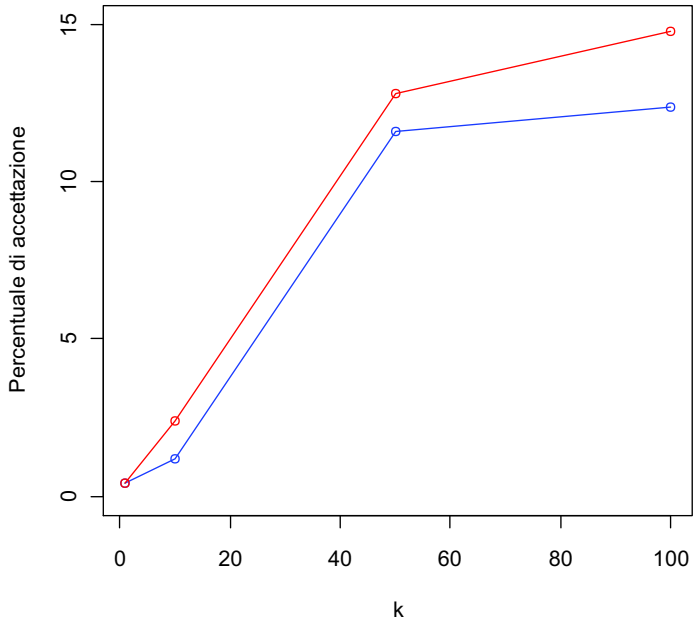

Grafico 7. Percentuali di accettazione del test (1) per un campione di numerosità campionaria pari a 500, al variare del numero k di dati estratti dalla singola distribuzione e del parametro α, assunto prima pari a zero e poi a 0.05. Mentre la linea blu rappresenta le percentuali di accettazione nel caso in cui il parametro di errore valga zero, quella rossa rappresenta le percentuali se il suddetto parametro vale 0.05.

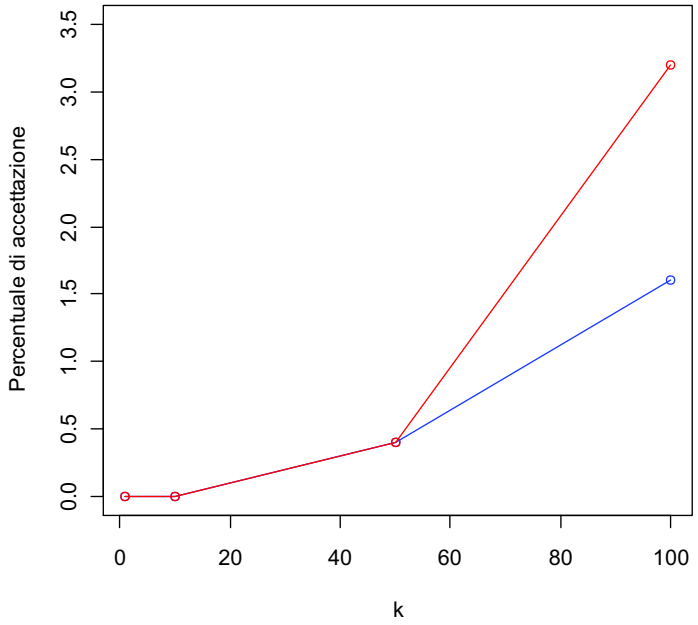

Grafico 8. Percentuali di accettazione del test (1) per un campione di numerosità campionaria pari a 1000, al variare del numero k di dati estratti dalla singola distribuzione e del parametro α, assunto prima pari a zero e poi a 0.05. Mentre la linea blu rappresenta le percentuali di accettazione nel caso in cui il parametro di errore valga zero, quella rossa rappresenta le percentuali se il suddetto parametro vale 0.05.

Anche in questo caso si nota che con l'aggiunta del parametro di errore, se aumentano i dati estratti dalla singola distribuzione, aumenta anche la percentuale di accettazione. Le differenze più marcate tra distribuzione "perturbata" e "non perturbata" si hanno per n = 250, 500. Nel caso in cui n = 1000 le differenze (anche se non troppo marcate) tra le due distribuzioni si hanno solamente quando il numero di dati estratti è pari a 100. Dai tre grafici si nota anche che con l'aggiunta di un parametro di errore si hanno delle percentuali di accettazione piuttosto basse, mai oltre il 30% al variare del numero k di dati casuali estratti dalla singola distribuzione e della numerosità campionaria n.

2.3. Modello GARCH.

Per vedere se le percentuali di accettazione aumentano o meno, si è provato ad utilizzare una funzione che restituisse i valori di un modello GARCH(1,1). Un modello autoregressivo a eteroschedasticità condizionata o modello GARCH (in inglese *Generalized AutoRegressive Conditional Heteroskedasticity*) è un modello utilizzato soprattutto nell'analisi delle serie storiche. In particolare le equazioni descriventi il modello sono le seguenti:

a. un'equazione descrivente la media condizionata al tempo t:

$$r_t = \upsilon_t + a_t \text{ , in cui:}$$

- $\upsilon_t = E[r_t / I_{t\ 1}]$,dove $I_{t\ 1}$ rappresenta l'informazione disponibile al tempo (t-1);

- $a_t = \sigma_t * \nu_t$, dove $\nu_t \sim N(0,1)\ \ \forall$ t;

b. un'equazione descrivente la varianza condizionata al tempo t:

$$\sigma_t^2 = \alpha_0 + \alpha_1 * (r_{t\text{-}1} - \upsilon_{t\text{-}1})^2 + \beta_1 * \sigma_{t\text{-}1}^2 .$$

Si è provato anche a verificare cosa accade al variare della numerosità del campione ottenuto dal GARCH(1,1). In particolare si è verificato cosa succede per le numerosità n = 250, 500, 1000, ripetendo tale procedura per cento volte. I parametri del modello sono stati fissati ai seguenti valori: υ_t pari a zero, α_0 pari a 0.05, α_1 pari a 0.3 ed infine β_1 pari a 0.5.

Nel Grafico 9 sono riportate le percentuali di accettazione del test che verifica l'aderenza dei dati generati da un GARCH(1,1) alla legge di Benford al variare della numerosità campionaria n.

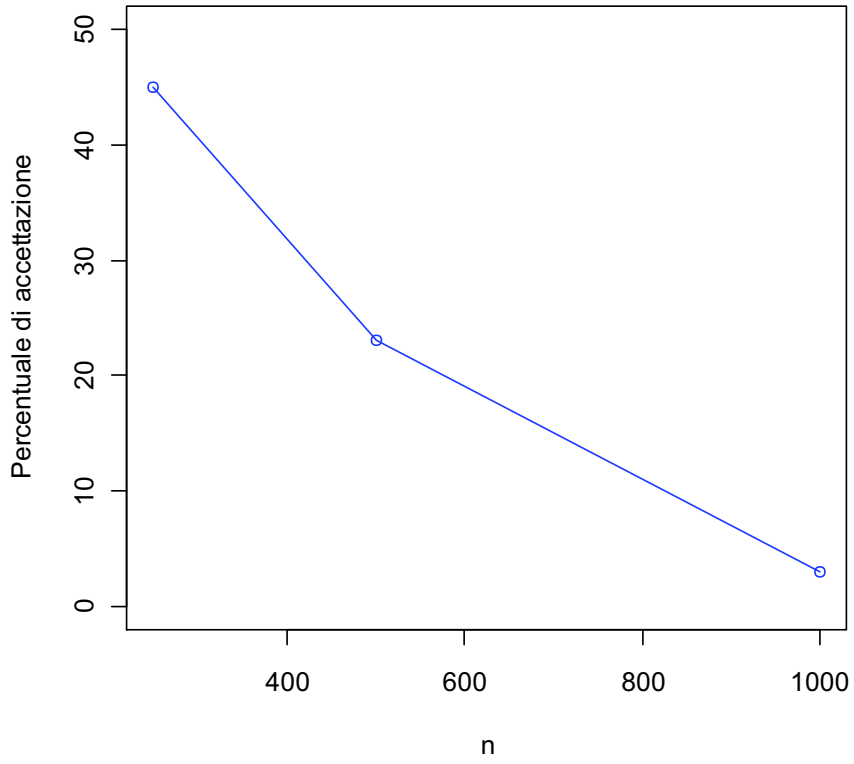

Grafico 9. Percentuali di accettazione del test (1) al variare della numerosità campionaria n per dati generati a partire da un modello GARCH(1,1).

Anche in questo caso si hanno delle percentuali di accettazione che diminuiscono all'aumentare della numerosità campionaria. Mentre con 1000 dati tale percentuale è piuttosto bassa (è infatti prossima allo zero), con 500 dati arriva al 25%, per poi arrivare quasi al 45% con 250 dati. Quindi anche con un campione di dati generati da un GARCH(1,1) pare che questi non rispettino la legge di Benford.

Si è provato a verificare se i risultati cambiassero se si fosse aggiunto un parametro α di errore pari a 0.05. Nel Grafico 10 sono rappresentate le percentuali di accettazione della distribuzione di dati con e senza il parametro di errore.

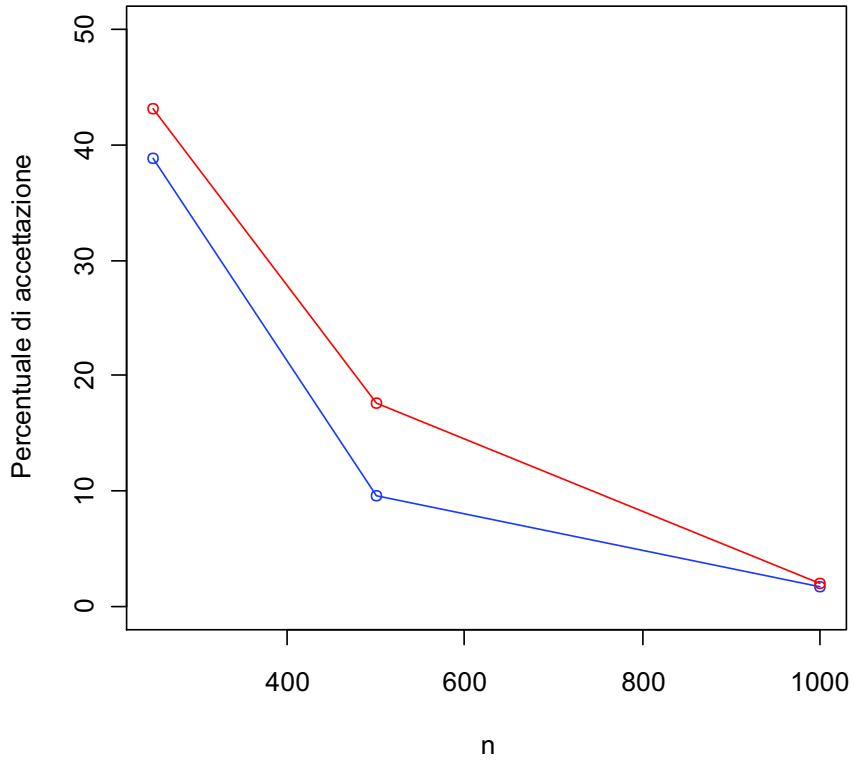

Grafico 10. Percentuali di accettazione del test (1) per dati generati a partire da un modello GARCH(1,1) al variare della numerosità campionaria n e del parametro di errore, assunto prima pari a zero e poi a 0.05. Mentre la linea blu è la percentuale di accettazione nel caso in cui il suddetto parametro valga zero, la linea rossa rappresenta la percentuale se il parametro vale 0.05.

Anche nel Grafico 10 si può vedere come la "distribuzione perturbata" abbia percentuali di accettazione sempre superiori a quelle della "distribuzione non perturbata". Tuttavia anche con l'aggiunta di un parametro di errore sembra che i dati non seguano la legge di Benford e ciò appare evidente soprattutto per le numerosità campionarie più elevate (n = 500, 1000).

3. Analisi di serie reali.

L'analisi delle serie reali si può suddividere in due parti.

Una prima parte nella quale è stato riportato quanto verificato da Benford. Analizzando otto dei ventuno elementi da lui considerati si è controllato quale valore del parametro α sia necessario per ottenere l'aderenza nel caso questa non fosse presente. Si analizzerà anche una successione numerica mostrando come, considerando la distribuzione delle frequenze relative della prima cifra significativa, essa sia molto simile a quella "teorica" della legge di Benford. Sarà effettuato anche il test (1) come ulteriore conferma dell'aderenza della successione considerata alla suddetta legge.

Nella seconda parte saranno considerati i rendimenti logaritmici giornalieri, settimanali e mensili di tre differenti indici azionari per lo stesso periodo di tempo. Dopo averne estratte la prima cifra significativa si verificherà se la distribuzione delle frequenze relative segua o meno la legge di Benford tramite il test (1). Se ciò non avviene si controllerà quale valore del parametro α sia necessario per poter ottenere il rispetto della suddetta legge.

3.1. Alcune analisi empiriche.

Inizialmente si è verificato se la distribuzione della prima cifra significativa della successione di Fibonacci seguisse o meno la legge di Benford. La suddetta successione si può ottenere assegnando i valori dei primi due termini F_0 (di solito preso pari a 0) e F_1 (di solito preso pari a 1) e chiedendo che ogni successivo sia $F_n = F_{n-1}+F_{n-2}$, con n>1. Per calcolare l'n-esimo termine della successione si utilizza la seguente equazione:

$$F_n = \left[\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n\right] = \left[\frac{\phi^n}{\sqrt{5}}\right]$$

dove:

a) ϕ è il cosiddetto rapporto aureo, pari a $\frac{1+\sqrt{5}}{2} = 1.618$;

b) [x] è la funzione che approssima x al suo intero più vicino, chiamata nint(x).

Dopo aver generato i primi duecento numeri di tale successione si sono ottenute le frequenze della prima cifra significativa, ed i risultati sono riportati in Tabella 3.

Prima cifra significativa	Frequenza relativa	Frequenza "teorica" legge di Benford
1	29.5%	30.1%
2	17.5%	17.6%
3	12.5%	12.5%
4	9.5%	9.7%
5	8.5%	7.9%
6	7%	6.7%
7	6%	5.8%
8	5%	5.1%
9	4.5%	4.6%

Tabella 3. Confronto delle frequenze relative dei primi duecento numeri della sequenza di Fibonacci con le frequenze "teoriche" ottenute dalla legge di Benford.

Dalla Tabella 3 si nota come le frequenze "teoriche" e quelle dei primi 200 numeri della successione di Fibonacci siano molto simili tra di loro. Come ulteriore conferma dell'aderenza dei primi 200 numeri della suddetta successione alla legge di Benford, si è calcolato il valore della statistica (1) che è risultato pari a:

Valore statistica (1) = 0.1687.

Se lo si confronta con il quantile di una distribuzione chi-quadrato con otto gradi di libertà (distribuzione asintotica) ottenuto ad un livello di significatività del 95%, pari a 15.5073, si nota che non si può rifiutare l'ipotesi nulla di uguaglianza della vera distribuzione dei primi 200 numeri della successione alla distribuzione "teorica" della legge di Benford.

Dopo aver applicato la suddetta legge alla successione di Fibonacci si è ripetuto ciò che aveva effettuato Benford nel testare la validità di quanto notato da Newcomb anni prima. In Tabella 4 sono riportate le frequenze ottenute dallo studioso nel caso di otto elementi sui ventuno da lui analizzati; è stato inoltre applicato il test (1) per verificare se le percentuali ottenute seguissero o meno la legge di Benford. Nel caso ciò non accadesse è stato riportato il valore del parametro α necessario per poter accettare l'ipotesi nulla del test (1).

	Tipologia dati	1	2	3	4	5	6	7	8	9	Numerosità campionaria	P-value test (1)	Valore α necessario per accettare H0
A	Rivers,Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335	0.76	0
B	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703	0.99	0
C	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690	0.90	0
D	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800	~0	0.20
E	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91	0.027	0.15
F	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741	0.048	0.04
G	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707	0.71	0
H	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418	0.478	0

Tabella 4. Frequenze relative ottenute da Benford nell'analisi di otto dei ventuno elementi da lui analizzati. Nelle ultime tre colonne sono riportate anche rispettivamente: a) la numerosità del campione considerato; b) il p-value ottenuto applicando la statistica (1) alle frequenze ottenute da Benford; c) il valore del parametro di errore necessario per poter accettare l'ipotesi nulla del test (1) ad un livello di significatività del 5%. Le tipologie di dati indicate nelle varie righe sono rispettivamente: A) "Rivers,Area" è la lunghezza misurata in miglia di alcuni fiumi americani; B) "Pressure" è la pressione di alcuni gas; C)"H.P. Lost" è il calore specifico perso da alcune sostanze chimiche in particolari condizioni; D) "Mol. Wgt." è il peso molecolare di alcuni composti chimici; E)"Atomic Wgt." è il peso atomico di alcuni elementi chimici;F) "Cost Data" è la lunghezza in miglia delle coste americane; G) "X-Ray Volts" è il potenziale elettrico prodotto da diversi raggi x; H) "Death Rate" sono gli indici di mortalità di alcuni Stati americani.

Osservando la Tabella 4 si nota in primis come gli elementi utilizzati da Benford fossero delle nature più differenti, spaziando dalle pressioni di alcuni gas, alla lunghezza delle coste americane per arrivare all'indice di mortalità. Si può accettare l'ipotesi nulla della statistica (1) in cinque casi su otto ed in quei casi in cui non la si accetta, il valore del parametro α è al massimo di 0.20. C'è poi da notare che l'accettazione o il rifiuto dell'ipotesi nulla del test (1) non dipende dalla numerosità del campione considerato: infatti così come si accetta la nulla con più di 700 dati, la si può rifiutare con sole 91 osservazioni.

3.2. Analisi dei rendimenti di tre indici azionari.

La seconda parte dell'analisi di serie reali si è concentrata sullo studio dell'aderenza alla legge di Benford dei rendimenti logaritmici di tre differenti indici azionari: il FTSE 100, il Nikkei 225 ed infine lo SP 500. Per i tre indici si sono calcolati i rendimenti logaritmici a partire dai loro prezzi di chiusura giornalieri, settimanali e mensili. Il periodo al quale si riferiscono i suddetti prezzi va dall'1 Gennaio 1990 fino all'11 Giugno 2010. In Tabella 5 per i tre indici sono presentate due statistiche descrittive di base: l'indice di asimmetria e l'eccesso di curtosi, ossia la differenza tra la curtosi del titolo e quella della normale standard, pari a 3.

Indice	Frequenza di osservazione	Asimmetria	Eccesso di curtosi
FTSE 100	Giornalieri	-0.089	6.413
	Settimanali	-0.847	3.809
	Mensili	-0.561	0.528
NIKKEI 225	Giornalieri	-0.019	5.190
	Settimanali	0.705	3.210
	Mensili	0.439	0.904
SP 500	Giornalieri	-0.202	8.966
	Settimanali	-0.769	3.169
	Mensili	-0.849	1.711

Tabella 5. Indice di asimmetria ed eccesso di curtosi per i rendimenti logaritmici dei tre indici azionari al variare della frequenza di osservazione.

Le due statistiche descrittive mostrate in Tabella 5 vengono qui presentate per vedere se possa esistere una loro relazione con l'accettazione o l'eventuale rifiuto dell'ipotesi nulla della statistica (1). E' da notare che al diminuire della frequenza di osservazione (ossia passando dai rendimenti giornalieri a quelli settimanali ed infine a quelli mensili) diminuisce anche l'eccesso di curtosi. Dopo aver visto le statistiche descrittive, in Tabella 6 sono presentati: i valori del test (1) applicato ai vari rendimenti, se si accetta o meno l'ipotesi nulla del suddetto test commettendo un errore del 5% ed il valore del parametro α che serve per accettare l'ipotesi nulla del test.

Indice	FTSE 100			NIKKEI 225			SP 500		
Frequenza	Giorno	Settimana	Mese	Giorno	Settimana	Mese	Giorno	Settimana	Mese
Valore statistica (1)	102.047	42.499	17.368	78.387	43.248	17.012	41.747	33.784	7.089
Si accetta H0 (errore 5%)	No	No	No	No	No	No	No	No	Sì
Valore di α per accettare H0	0.090	0.040	0.015	0.065	0.040	0.015	0.040	0.030	0

Tabella 6. Valori della statistica (1) per i rendimenti dei tre indici al variare della frequenza. Per stabilire se si accetta H0 si è confrontato il valore della suddetta statistica con il quantile di una distribuzione chi-quadrato con otto gradi di

libertà (distribuzione asintotica della statistica (1)) ottenuto ad un livello di significatività del 95% e pari a 15.5073. Nell'ultima riga si ha poi il valore del parametro α necessario per poter accettare l'ipotesi nulla.

Dalla Tabella 6 appare come al diminuire della frequenza di osservazione diminuiscono anche il valore della statistica (1) calcolato a partire dai rendimenti logaritmici e il valore del termine di errore α necessario per poter accettare l'ipotesi nulla del test (1). Senza l'ausilio di α si può accettare l'ipotesi nulla di aderenza dei rendimenti alla legge di Benford solo per i rendimenti logaritmici mensili dello SP 500.

Confrontando quanto ottenuto in Tabella 6 con quanto ottenuto in Tabella 5 si nota che un minor eccesso di curtosi implica un valore della statistica (1) più basso e quindi una maggior probabilità di accettare l'ipotesi nulla del suddetto test.

Il fatto che al diminuire della frequenza di osservazione diminuisce anche il valore ottenuto con la statistica (1) può essere dovuto principalmente a due motivi. Il primo è che i rendimenti mensili hanno meno osservazioni rispetto a quelli settimanali, che a loro volta ne hanno meno di quelli giornalieri; con il diminuire della numerosità campionaria anche il test (1) perde di potenza ed è forse per questo che con i rendimenti mensili si accetta l'ipotesi nulla con l'ausilio di un α praticamente pari a zero per tutti e tre gli indici. La seconda causa si può ricercare nel fatto che i prezzi mensili (ed in minor misura anche i settimanali) e quindi anche i rendimenti mensili di un indice "assorbono" al proprio interno tutte quelle piccole variazioni che invece influenzano i giornalieri. Quindi, probabilmente, il minor valore della statistica (1) è dovuto al fatto che i rendimenti giornalieri sono maggiormente soggetti a variazioni nell'arco di breve tempo; queste si traducono in una maggiore variabilità delle prime cifre significative e ciò provoca una maggiore probabilità di ottenere una prima cifra significativa più elevata rispetto ai rendimenti settimanali e mensili. Per questi ultimi le prime cifre significative si possono quindi considerare più stabili.

4. Conclusioni.

Perché dei dati casuali dovrebbero rispettare la legge di Benford? La suddetta legge non sarebbe valida se si prendessero in considerazione tutti i numeri compresi tra uno ed infinto; in questo caso infatti ogni potenziale prima cifra sarebbe equamente rappresentata. Però, dato che i numeri con i quali si ha a che fare nella realtà quotidiana hanno sempre un valore finito e vengono spesso generati in ordine crescente, si può comprendere come mai quelli che iniziano con una cifra di basso valore abbiano una maggiore probabilità di essere generati rispetto a quelli che iniziano con una cifra di valore più elevato.

Da quanto ottenuto in questo elaborato si può dedurre che non si può affermare a priori quando un insieme di dati casuali segue o meno la legge di Benford: in particolare, quando sono state estratte le osservazioni da una singola distribuzione, si è visto come passando da una distribuzione generatrice all'altra i risultati cambiassero di molto. Nel caso invece dei dati reali si è solamente rilevato che al diminuire della frequenza di rilevazione dei rendimenti diminuisce il valore della statistica (1), ma ciò non sempre porta ad accettare l'ipotesi nulla di aderenza della distribuzione dei dati casuali alla distribuzione "teorica" della legge di Benford.

Quindi non si è riusciti a delineare i confini entro i quali dei dati casuali (reali o simulati) rispettano la legge di Benford, ma si sono ottenuti dei casi notevoli in cui ciò accade.

- Bibliografia.

- Benford, F. (1938), The Law of Anomalous Numbers, *Proceedings of the American Philosophical Society*, Vol. 78, No.4, pp. 551-572;
- Corazza, M., Ellero, A., Zorzi, A. (2010), Checking Financial Markets via Benford's Law: the S&P 500 Case, *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pp. 93-102;
- Fewster, R. M. (2009), A Simple Explanation of the Benford's Law, *The American Statistician*, Vol. 63, No. 1, pp. 26-32;
- Hill, T. P. (1995), A statistical derivation of the Significant-Digit Law, *Statistical Science,* Vol.10, No.4, pp. 354-363;
- Ley, E. (1996), On the Peculiar Distribution of the U.S. Stock Indexes' Digits, *The American Statistician*, Vol. 50, No. 4, pp. 311-314;
- Newcomb, S. (1881), Note on the Frequency of Use of the Different Digits in Natural Numbers, *American Journal of Mathematics,* Vol. 4, No. 1, pp. 39-40;
- Nigrini, M. (1996), A Taxpayer Compliance Application of Benford's Law, *Journal of the American. Taxation Association*, Vol. 18, No. 1, pp. 72-92;
- Pinkham, R. S. (1961), On the Distribution of First Significant Digits, *Annals of Mathematical Statistics*, Vol. 32, No. 4, pp. 1223-1230;
- Varian, H. R. (1972), Benford's Law, *The American Statistician*, Vol. 26, pp. 65-66.