· Università degli Studi di Padova

FACOLTÀ DI SCIENZE STATISTICHE
Corso di Laurea Triennale in Statistica e Tecnologie Informatiche

Modello di change-point per dati esponenziali

Relatore:
GUIDO MASAROTTO

Laureando:
EMANUELE GIORGI

Anno Accademico: 2009 - 2010

Al mio unico pilastro:
Mamma

"...dai diamanti non nasce niente dal letame nascono i fior..."

Fabrizio De Andrè

Indice

Introduzione

Il problema affrontato nella tesi si colloca nell'ambito del Controllo Statistico di Processo (d'ora in avanti SPC), quel ramo della statistica che rivolge il suo interesse in particolar modo ai processi di produzione industriale (e non solo) o meglio ad una o più particolari caratteristiche misurabili, secondo una qualche scala, del processo. Ci si preoccupa di studiare, qualora non sia già nota, la distribuzione probabilistica di tale caratteristica quando il processo è in controllo, ossia quando la variabilità è quella naturale del processo e la produzione di un'unità non conforme è dovuta solo ed esclusivamente al caso, ossia alle cosiddette cause speciali. Parleremo di cause assegnabili (ad esempio usura di un macchinario, nuovo operatore, nuovo macchinario,...) quando un cambiamento nella media e nella varianza o una qualsiasi altra forma di deviazione dalla distribuzione in controllo sono dovuti a fonti di variabilità che possono essere rilevate e pertanto rimosse. La qualità infatti è inversamente proporzionale alla variabilità e nostro obiettivo è ridurre questa il più possibile al fine di ricondurci a quella minima, naturale del processo. Lo statistico in tutto questo è supportato dalle cosiddette carte di controllo, uno strumento statistico atto a sorvegliare la caratteristica del processo di interesse e che segnala più o meno velocemente (a seconda di quale carta di controllo utilizziamo e come la disegniamo) uno stato di fuori controllo. Come vedremo più avanti un punto importante sarà fissare la probabilità con la quale la nostra carta segnala un falso allarme. Nell'applicazione delle carte di controllo si distinguono due stadi, che sono gli studi di Fase I e Fase II: nel primo momento si dispone di un insieme retrospettivo di dati e si cerca di capire se questi provengono da un processo in controllo, se questo non è verificato bisognerà eliminare dai dati quelli in cui si registra il cambiamento investigando sulla natura delle cause assegnabili che lo hanno generato; successivamente ci si occupa del monitoraggio del processo e se le fonti di maggior variabilità sono state rimosse durante la Fase I in genere (ma non è detto) avremo a che fare con cambiamenti di minore entità.

Monitorare eventi rari

La nostra attenzione sarà rivolta al problema del monitoraggio di eventi rari, o meglio degli intervalli di tempo che intercorrono nel manifestarsi di due eventi rari consecutivi. Per evento raro si intende un evento che in un dato intervallo temporale ha una probabilità molto bassa di verificarsi. Ovviamente tutto è relativo all'intervallo di tempo che si considera poichè la nascita di un bambino in

un dato ospedale non è certo un evento raro se si ha come tempo di osservazione 30 (o anche meno) giorni ma sicuramente lo è se si considerano i successivi 30 secondi. Numerosi sono gli esempi di eventi rari nei quali è necessario il monitoraggio: i segnali di errore di un software di sistema, le pulsazioni lungo una fibra nervosa, diffetti di produzione, malattie rare in una data regione, incidenti sul lavoro, terremoti, disastri industriali,...
Prendendo a rifermento i processi produttivi industriali (che in particolar modo nelle industrie manifatturiere hanno conosciuto una forte innovazione tecnologica nel secolo scorso) sono caratterizzati da una bassa difettosità che rende più difficoltosa l'applicazione di carte per attributi, quali ad esempio le carte c ed u per il conteggio dei diffetti con le quali potremmo avere una serie di valori pari a 0 e un fuori controllo per un qualsiasi valore diverso da 0. E' chiaro pertanto come queste carte possano risultare poco informative. In un processo di Poisson[1] che sia omogeneo (l'intensità del processo è costante) da poter così rispettare le proprietà di uniformità (la probabilità di avere un certo numero di eventi dipende solo dall'ampiezza dell'intervallo temporale e non dalla sua posizione sull'asse dei tempi), di isolatezza (abbiamo a che fare con un eventi rari, dunque in un intervallo di tempo sufficientemente piccolo è ammesso al massimo il verificarsi di un solo evento) e di indipendenza (il numero di eventi occorsi in intervalli temporali disgiunti sono realizzazioni di variabili casuali indipendenti tra loro e da quanto accaduto prima) i tempi intercorrenti tra la rilevazione di due difetti consecutivi hanno distribuzione esponenziale con media il reciproco dell'intensità del processo e sono indipendenti tra di loro. L'assunzione distributiva che d'ora in avanti faremo è che:

$$Y_i \sim Exp(\mu)$$

dove Y_i è la i-esima v.a. che determina il tempo intercorrente nella rilevazione di due difetti consecutivi (parleremo in seguito di "tempo tra la rilevazione di unità difettose" per avere un riferimento reale immediato, ma come abbiamo già detto la natura di tale evento può essere qualsiasi). Ci concentreremo in particolar modo su un modello di change-point che ha questa formulazione:

$$\begin{cases} Y_i \sim Exp(\mu_1) & i = 1, \dots, \tau - 1 \\ Y_i \sim Exp(\mu_2) & i = \tau, \dots, n \end{cases}$$

$$\mu_1 > 0, \quad \mu_2 > 0, \quad \mu_1 \neq \mu_2, \quad \tau \in \{2, \dots, n+1\}$$

si ha dunque dal punto di svolta τ in poi un cambiamento della media del nostro processo (ed anche nella varianza data la natura esponenziale). Vedremo l'utilità di questo modello quando i parametri di interesse sono tutti ignoti (il caso più frequente nella realtà) facendo uso del metodo di massima verosimiglianza. La carta di controllo basata sul modello di change-point (che chiameremo in seguito per brevità carta GLR, *Generalized Likelihood Ratio*) verrà presentata nel Capitolo 2 ripercorrendo l'esposizione di Hawkins (2003 e 2005) nel caso di dati distribuiti normalmente al fine di evidenziare l'analogia dei risultati ottenuti nonostante l'ipotesi distributiva sui dati sia diversa. Il problema che

[1] Viene qui considerata la definizione "temporale" del processo di Poisson, nonostante anche la definizione "spaziale" sarebbe appropriata ma dato che si parlerà di tempi di attesa viene considerata solo la prima.

affronteremo sarà in sostanza l'adattamento della formulazione del modello di change-point al caso esponenziale.

Capitolo 1

Assunti ed una prima soluzione

Nella prima parte di questo capitolo discuteremo come verificare gli assunti necessari per l'applicazione delle carte di controllo presentate successivamente. Nella seconda parte viene proposta una prima soluzione del problema su come monitorare il processo senza far riferimento al modello di change-point e utilizzando le carte di controllo Shewart e CUSUM.

1.1 Verifica dell'assunto distributivo e di indipendenza

Se la distribuzione del numero di difetti non segue la legge di probabilità di Poisson, o meglio non possiamo riferirci ad un processo di Poisson allora le Y_i non hanno distribuzione esponenziale. E' necessaria dunque una verifica attraverso opportuni strumenti grafici e verifiche di ipotesi. Infine l'ipotesi d'indipendenza sarà un altro assunto di cui andrà verificata la validità.

Strumenti grafici e test di esponenzialità

Disponendo di un campione di numerosità n verifichiamo se l'assunto distributivo esponenziale ha senso mediante gli usuali strumenti grafici per v.a. continue: istogrammi, funzione di ripartizione empirica, boxplot, Q-Q plot, P-P plot,... Ovviamente dobbiamo fare i conti con la numerosità campionaria: tanto più grande è la dimensione del nostro campione maggiore è l'informazione (escludendo situazioni particolari: dati mancanti, dati censurati,etc...) che abbiamo ma maggiore è anche il costo. Per conoscere bene una qualsiasi v.a. dobbiamo tener presente anche i momenti che la caratterizzano, più precisamente la

media, la varianza, l'indice di asimmetria e l'indice di curtosi che per una v.a. $Y \sim Exp(\mu)$ sono rispettivamente:

$$E[Y] = \mu$$

$$E\left[(Y-\mu)^2\right] = \mu^2$$

$$E\left[\left(\frac{Y-\mu}{\mu}\right)^3\right] = 2$$

$$E\left[\left(\frac{Y-\mu}{\mu}\right)^4\right] = 6$$

E' utile disporre di una stima campionaria dell'indice di asimmetria e possiamo ad esempio calcolare la seguente quantità:

$$\hat{\gamma} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{3/2}} \tag{1.1}$$

Lo stimatore $\hat{\gamma}$ è distorto, ossia il suo valore atteso per un dato n è diverso da 2 ma è consistente. Si possono utilizzare altri stimatori proposti in letteratura che risultano essere non distorti. Dovendo verificare l'ipotesi $H_0 : \gamma = 2$ un metodo è quello di ottenere via Monte Carlo una stima della distribuzione campionaria dello stimatore in (1.1) e ottenere così il valore soglia con il quale confrontare il valore della statistica test calcolata sui dati. Un analogo discorso può essere fatto anche per quanto riguarda la curtosi. Riportiamo in Figura 1.1 alcuni Q-Q plot per rendere più chiaro che tipo di discostamento sistematico[1] ci aspettiamo quando i dati non provengono da una v.a. esponenziale. Si osserva che quando i dati provengono da una distribuzione meno asimmetrica dell'esponenziale si ha una "u" con la concavità verso il basso, mentre se è più asimmetrica la "u" rivolge la concavità verso l'alto. Si propone ora un test di esponenzialità che va sotto il nome di Shapiro-Wilk, che saggia il seguente sistema di ipotesi:

$$H_0 : f(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}, \quad H_1 : f(x) \neq \frac{1}{\mu}e^{-\frac{x}{\mu}}, \quad x > 0, \quad \mu > 0$$

Dove $f(x)$ è la funzione di densità della v.a. X. La statistica test è la seguente:

$$T_{S.W.} = \frac{n(\overline{X} - X_{min})^2}{[(n-1)S]^2}$$

$\overline{X}$ è la media campionaria delle osservazioni, X_{min} è il valore minimo campionario ed S è la radice quadrata della varianza campionaria corretta (la statistica test sotto H_0 non dipende da μ).

[1] La retta nei grafici interpola il primo quartile ed il nono decile in maniera tale da cogliere meglio l'asimmetria dei dati.

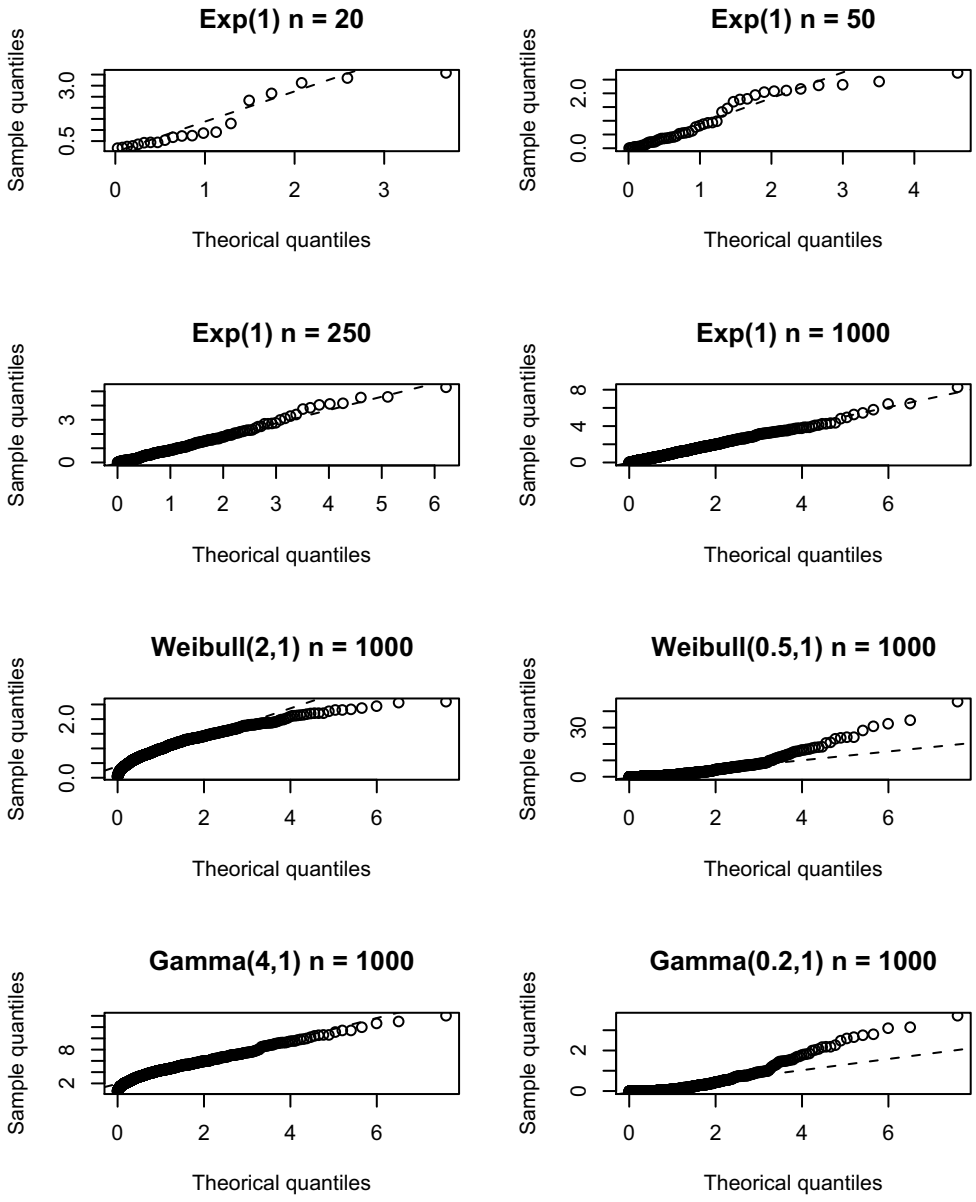

Figura 1.1: Una serie di Q-Q plot (sopra ogni riquadro viene indicata la distribuzione dalla quale i dati sono sati generati; nel caso della Weibull e della Gamma viene indicato tra parentesi prima il parametro di forma e poi quello di scala)

Distribuzioni probabilistiche per tempi di attesa

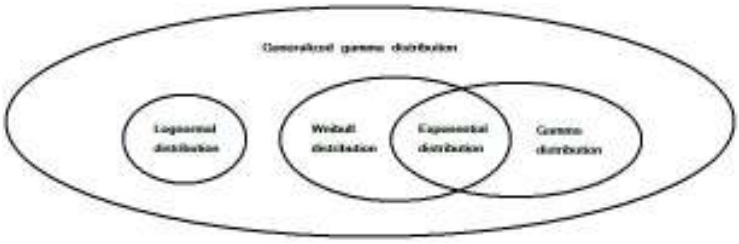

Figura 1.2: La v.a. Gamma generalizzata ed alcuni casi notevoli

Consideriamo ora una certa categoria di v.a. che si possono incontrare quando si ha a che fare con dati che rappresentano tempi di attesa. La Gamma generalizzata è una v.a. continua la cui funzione di densità è caratterizzata da tre parametri:

$$f(x) = \frac{\beta}{\Gamma(k)\theta}\left(\frac{x}{\theta}\right)^{k\beta-1} e^{-(\frac{x}{\theta})^{\beta}}$$

$\theta > 0$ è il parametro di scala, mentre $k > 0$ e $\beta > 0$ sono i parametri di forma (il supporto della v.a. è $x > 0$). Considerando delle particolari riparametrizzazioni, che qui non specifichiamo[2], e particolari valori per queste si ricade in uno dei tre sottotinsiemi (il caso Lognormale è un caso limite per $k \to \infty$) che si vedono nella Figura 1.2. Molti altri potrebbero essere i sottoinsiemi da considerare, ma questi sono sicuramente i più diffusi nelle applicazioni e che spesse volte si "confondono" con il modello esponenziale. Dai grafici in Figura 1.3 possiamo intuire come, anche con un campione di non modesta numerosità, gli strumenti diagnostici ci possano portare all'accettazione dell'assunto distributivo esponenziale quando siamo vicini all'intersezione tra gli insiemi Gamma e Weibull e ci troviamo in realtà in uno dei due oppure ancora abbiamo una Lognormale con un parametro di forma minore di 1 ed un parametro di scala tali che la somiglianza distributiva con l'esponenziale è forte. Possiamo dire che la probabilità di commettere un errore del secondo tipo in questi casi è molto alta, ma in realtà l'approssimazione all'esponenziale è sensata ed anche esemplificatrice nell'interpretazione del fenomeno.

L'indipendenza

L'indipendenza dei tempi di attesa si ha se il processo di Poisson sottostante gode anch'esso della proprietà di indipendenza. Possiamo utilizzare strumenti grafici come il correlogramma ed i lag-plot, test come quello di Ljung-Box per la verifica dell'assunto di incorrelazione, anche se con tali strumenti in realtà verifichiamo se i dati sono incorrelati (e come sappiamo l'indipendenza implica l'incorrelazione ma non viceversa). Esistono diversi metodi per la verifica dell'indipendenza stocastica (ossia l'indipendenza più forte e generale che possiamo concepire) che qui tralasciamo rimandando all'esempio dell'ultimo capitolo in cui viene utilizzato un test χ^2.

[2] per maggiori dettagli sulla Gamma generalizzata si veda il sito web *www.weibull.com* o il libro di testo *Statical Models and Methods for Lifetime Data*, Lawless (2nd ed.)

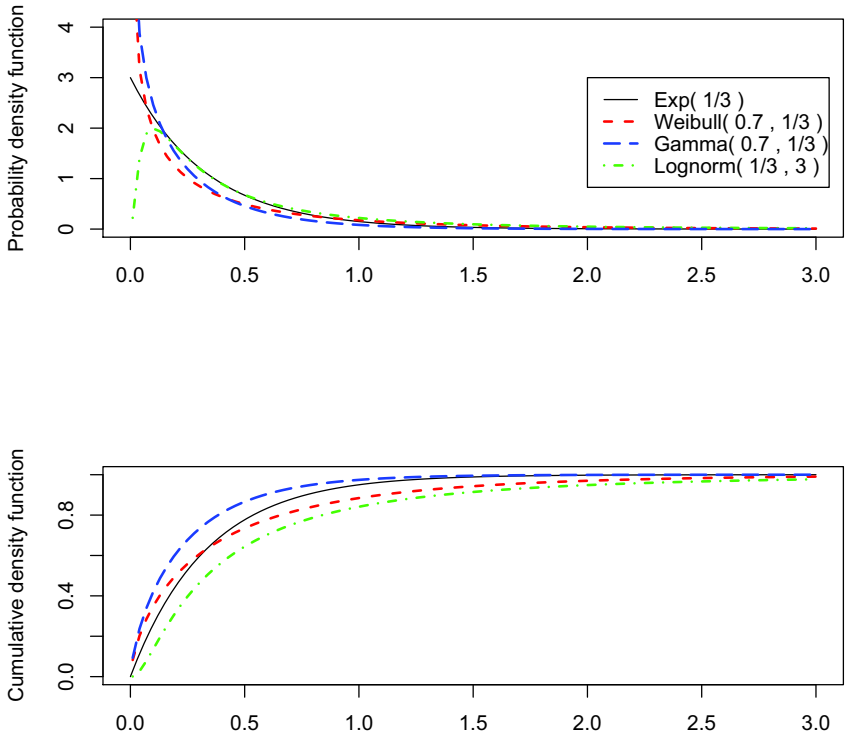

Figura 1.3: Alcuni casi in cui le distribuzioni hanno un andamento "quasi" esponenziale.

1.2 La carta Shewart per dati esponenziali

Data la difficoltà nell'applicazione della carta di controllo Shewart a dati esponenziali dovuta alla forte asimmetria di tale distribuzione, è stata proposta da Nelson una semplice trasformazione di potenza che ci fa passare dalla distribuzione esponenziale alla distribuzione di Weibull. Se $X \sim Exp(\lambda)$ ed operiamo la trasformazione $X^{\frac{1}{3.6}}$, si ha che:

$$P[X^{\frac{1}{3.6}} \leq x] = P[X \leq x^{3.6}] = 1 - e^{-\lambda x^{3.6}}$$

La funzione di densità è dunque (il supporto è dato da $x > 0$):

$$\begin{aligned} f(x) &= 3.6\lambda x^{3.6-1} e^{-\lambda x^{3.6}} \\ &= \frac{3.6}{\lambda^{-\frac{1}{3.6}}} \left(\frac{x}{\lambda^{-\frac{1}{3.6}}} \right)^{3.6-1} e^{-\left(\frac{x}{\lambda^{-\frac{1}{3.6}}} \right)^{3.6}} \end{aligned}$$

D'ora in avanti considereremo la riparametrizzazione $\delta = \lambda^{-\frac{1}{3.6}}$ ed in definitiva:

$$X^{\frac{1}{3.6}} \sim Weibull(3.6, \delta)$$

Dove 3.6 è il parametro di forma e δ è il parametro di scala. Il motivo di questa trasformazione è dato dal fatto che una distribuzione di Weibull con parametro di forma 3.6 è molto simile alla distribuzione normale. Per simile si intende il fatto che le due distribuzioni hanno l'indice di asimmetria e di curtosi standardizzati che differiscono di poco. Infatti dato il parametro di forma pari a 3.6 e di scala pari a δ la media e la deviazione standard della Weibull sono rispettivamente:

$$\begin{aligned} \mu &= \delta\Gamma\left(1+\frac{1}{3.6}\right) \\ \sigma &= \sqrt{\delta^2\Gamma\left(1+\frac{2}{3.6}\right)-\mu^2} \end{aligned}$$

Infine l'indice standardizzato di asimmetria e di curtosi risultano essere:

$$\begin{aligned} \gamma_1 &= \frac{\delta^3\Gamma\left(1+\frac{3}{3.6}\right)-3\mu\sigma^2-\mu^3}{\sigma^3} \\ \gamma_2 &= \frac{\delta^4\Gamma\left(1+\frac{4}{3.6}\right)-4\gamma_1\mu\sigma^3-6\mu^2\sigma^2-\mu^4}{\sigma^4} \end{aligned}$$

Effetuando alcuni semplici passaggi si può far vedere che sia γ_1 che γ_2 non dipendono da δ. Si ha così che l'indice di asimmetria è circa $5.6 \cdot 10^{-4}$ mentre la curtosi è circa 2.72, non molto lontani dai rispettivi valori di una v.a. Normale, ossia 0 e 3.

Come procedere dunque nell'utilizzo della carta Shewart è semplice: una volta che si dispone di un campione di tempi di attesa intercorsi tra due difetti consecutivi trasformiamo i valori elevandoli a $1/3.6$ e possiamo ora tarattare i dati come fossero normali e adottare il solito procedimento per dati normali nel disegno della carta Shewart e nell'applicazione della carta di controllo in Fase I ed in Fase II.
Noi vedremo la carta di controllo Shewart per monitorare cambiamenti nella media del processo (anche se sappiamo che, nel caso che trattiamo, un cambiamento nella media coinvolge direttamente un cambiamento anche nella varianza del processo). Supponiamo inoltre che ci siano due individui, uno che utilizza la carta Shewart e l'altro invece la carta di controllo GLR: ebbene al primo daremo un grande vantaggio, ossia disporrà del vero valore del parametro in controllo e dunque non sarà necessario uno studio preliminare di Fase I (questo fatto nella realtà è più che raro). Nonostante questo vantaggio molto più importanti saranno gli svantaggi a cui andrà incontro: come abbiamo già accennato un cambiamento nella media del processo coinvolge anche la deviazione standard dello stesso, dunque assumerà un ruolo cruciale lo studio dell'andamento delle

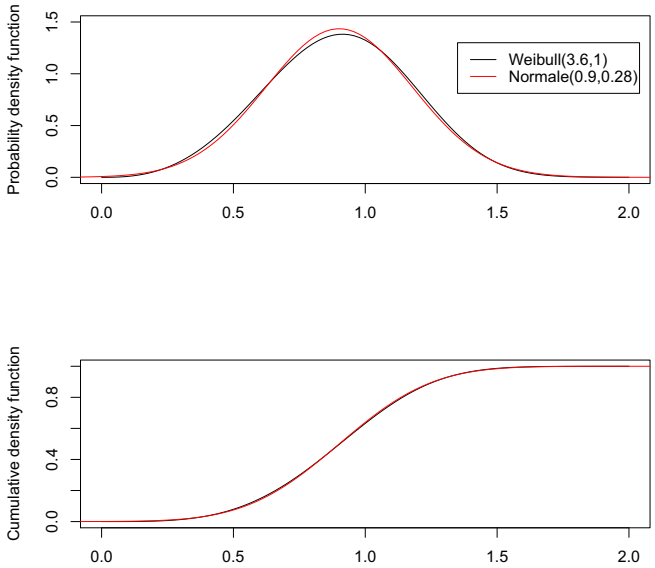

Figura 1.4: La distribuzione di Weibull con parametro di forma 3.6 e di scala 1 a confronto con una Normale che ha per media e deviazione standard le stesse della suddetta Weibull.

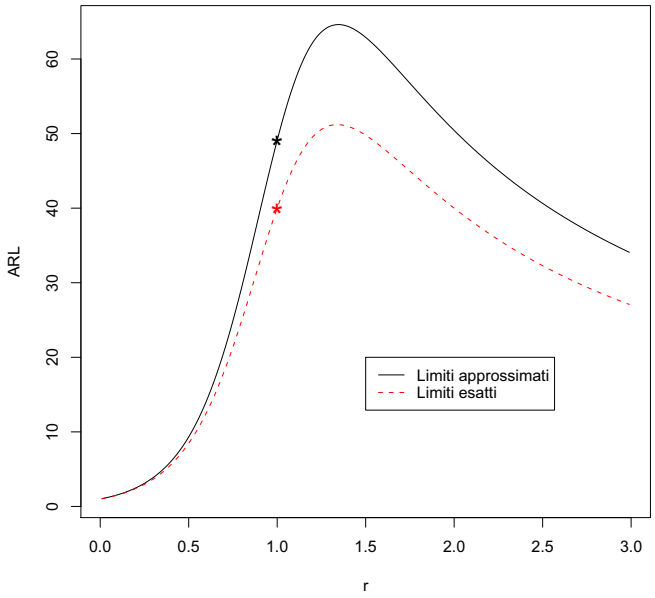

Figura 1.5: I profili dell'ARL per una carta di controllo Shewart per la media del processo; per limiti esatti si intendono i limiti calcolati a partire dai quantili della Weibull, per limiti approssimati si considera invece l'approssimazione alla Normale; la probabilità di commettere un errore del primo tipo è fissata ad $\alpha = 0.025$. Ricordiamo infine che $r =$ (media in controllo)/(media fuori controllo).

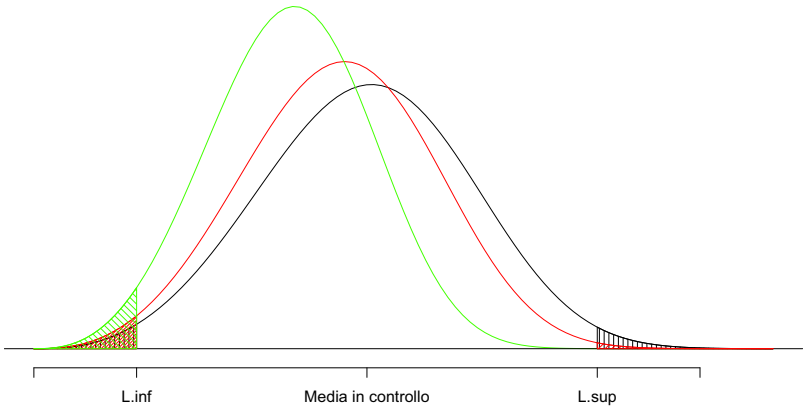

Figura 1.6: Tre Weibull a confronto: in nero il caso in controllo e le altre due sono casi di diminuzione della media del processo in controllo.

osservazioni (trasformate) entro i limiti di controllo[3], infatti ad una diminuzione della media del processo corrisponde una diminuzione della deviazione standard che rende il limite di controllo superiore eccessivamente alto mentre per quello inferiore dipende anche da quanto diminuisce la media (poichè i valori fluttueranno di meno su un livello medio più basso)[4]. Il legame esistente tra media e varianza determina un profilo dell'ARL particolare. Tenendo conto del legame tra media e varianza e di quanto detto prima si vede in Figura 1.5 come per alcune diminuzioni nella media del processo l'ARL fuori controllo è addirittura maggiore di quello in controllo; dalla stessa immagine emerge inoltre il fatto che, nonostante la forte somiglianza tra la Weibull (con le caratteristiche precedentemente definite) e la Normale, le differenze in termini di ARL sono notevoli.
In Figura 1.6 sono riportate tre Weibull aventi tutte e tre parametro di forma 3.6 e parametri di scala 1 (nero), 0.92 (rosso) e 0.84 (verde): la Weibull con parametro di scala pari ad 1 rappresenta il caso in controllo, quella con parametro di scala pari a 0.92 rappresenta invece il caso in cui l'ARL è massimo ed infine quella con parametro di scala pari a 0.84 a un ARL fuori controllo uguale a quello in controllo (se si confrontano le aree colorate tutto ciò è evidente). Possiamo dunque concludere che la carta Shewart per dati esponenziali

[3] Come sappiamo questo può portare ad una diminuzione dell'ARL in controllo.

[4] Nel caso di un aumento della media del processo, dunque anche della deviazione standard, le osservazioni non tendono più a rimanere entro i limiti di controllo come nel caso della diminuzione, ma la tendenza è opposta; questo di fatto può essere un vantaggio anche se in realtà è più importante rilevare in tempo un peggioramento del sistema produttivo (dovuto ad una diminuzione del tempio medio che intercorre tra due difetti) piuttosto che un miglioramento.

è poco efficiente nell'individuare peggioramenti nel sistema produttivo ed ha invece buone prestazioni nell'individuare miglioramenti del sistema produttivo.

1.3 La carta CUSUM per dati esponenziali (cenni)

Introduciamo brevemente la carta CUSUM bilaterale (in forma algoritmica) per dati esponenziali; abbiamo che il valore della statistica, che rileva le diminuzioni nella media, nell'istante i-esimo è dato da:

$$T_i = \min(0, T_{i-1} + x_i - k_T)$$

Mentre quello della statistica che rileva gli aumenti è dato da:

$$S_i = \max(0, S_{i-1} + x_i - k_S)$$

Dove k_T e k_S sono dei valori che derivano dal SPRT (*Sequential Probability Ratio Test*), e sono dati da (μ_1 è la media in controllo e μ_2 è la media fuori controllo):

$$k_J = \frac{\mu_1 \mu_2}{\mu_2 - \mu_1} \log \frac{\mu_2}{\mu_1}$$

Se $J = T$ allora $\mu_1 > \mu_2$, se $J = S$ allora $\mu_1 < \mu_2$. Viene lanciato un allarme ogni qualvolta:

$$S_i > h_S \quad \vee \quad T_i < -h_T$$

Dove h_S ed h_T sono i limiti di controllo che vengono calcolati una volta fissati l'ARL in controllo (chiamermo $H(0)$ l'ARL in controllo che si ha nella rilevazione di aumenti nella media in controllo ed $L(0)$ quello nelle diminuzioni) e il valore k_J per $J = S, T$. L'ARL globale è dato dalla seguente approssimazione (che sotto particolari condizioni risulta essere esatta):

$$ARL_0 \doteq \frac{H(0)L(0)}{H(0) + L(0)}$$

Per approfondimenti e chiarimenti sulla CUSUM per dati esponenziali si rimanda a Gan (1994).

Capitolo 2

Il modello di change-point

Le realizzazioni dei tempi di attesa, dato un punto di svolta τ ed un campione di ampiezza fissa pari ad n, possono essere modellate come segue:

$$\begin{cases} Y_i \sim Exp(\mu_1) & i = 1, \dots, \tau - 1 \\ Y_i \sim Exp(\mu_2) & i = \tau, \dots, n \end{cases}$$

$$\mu_1 > 0, \quad \mu_2 > 0, \quad \mu_1 \neq \mu_2, \quad \tau \in \{2, \dots, n+1\}$$

Il modello di change-point è stato formulato per la rilevazione di cambiamenti dovuti a cause assegnabili e che pertanto permangono finchè non c'è un intervento di aggiustamento del processo. La carta di controllo basata sul modello di change-point rientra nella categoria delle carte con memoria. Ipotizziamo inoltre che il punto di svolta sia unico, ossia se un cambiamento avviene nel processo questo è unico. Date le peculiarità della distribuzione esponenziale questo cambiamento interessa direttamente sia la media che la varianza in controllo.
Si precisa inoltre che in un contesto di Fase I noi disponiamo di un campione di tempi di attesa con numerosità n fissata e pertanto il punto di svolta τ può variare (nell'insieme dei valori interi) da 2 fino ad n se il cambiamento è occorso e se non abbiamo nessun cambiamento allora $\tau = n + 1$ che è equivalente alla scrittura $\tau > n$ poichè siamo in Fase I. Se siamo in Fase II invece e assumiamo che il processo è in controllo (ossia siamo sotto H_0) l'insieme di definizione di τ non è più limitato perchè n non è più costante.
Il modello di change-point, come vedremo avanti, è impiegato oltre che per la costruzione della statistica test anche come strumento per le stime di μ_1, μ_2 e τ.
Inoltre se $\mu_1 > \mu_2$ allora avremo un peggioramento del sistema produttivo poichè il tempo medio di attesa tra difetti consecutivi è diventato più piccolo mentre se $\mu_1 < \mu_2$ avremo un miglioramento del sistema produttivo.

2.1 Primo scenario: μ_1 e μ_2 noti

In questo primo scenario le carte di controllo più adatte sono la CUSUM e la EWMA per dati esponenziali poichè data la media in controllo μ_1 possiamo ottenere il disegno ottimo per la rilevazione del cambiamento della media in controllo al suo mutliplo μ_2 (ovviamente τ rimane l'unico parametro ignoto da stimare). In questo caso può pertanto essere sfruttata l'ottimalità della CUSUM o dell'EWMA con cambiamenti verso valori μ_2. Per ottenere maggiori dettagli su queste carte si veda Gan (1994) per la CUSUM e Gan (1998) per la EWMA.

2.2 Secondo scenario: μ_1 noto e μ_2 ignoto

E' raro conoscere il valore di μ_2 verso il quale avviene il cambiamento, pertanto il disegno della carta CUSUM viene fatto considerando un valore multiplo della media in controllo accettabile per il quale l'ARL fuori controllo sia minimo; infatti anche per valori di μ_2 che non sono esattamente uguali al valore fuori controllo per il quale si è disegnata la carta CUSUM, questa garantisce performance di poco distanti da quelle teoriche. Analogo il discorso per la carta EWMA. Può essere impiegata la carta GLR, sia per verificare se il cambiamento è occorso e stimare τ e μ_2.

Da dove vengono i valori noti dei parametri? Il valore μ_1 che abbiamo precedentemente definito come noto è in realtà una stima che proviene da uno studio di Fase I, dove stiamiamo μ_1 usando i metodi statistici per campioni di ampiezza fissa, assicurandoci che i dati provengono da un processo in controllo. Ma si è visto che stime anche con piccoli errori, essendo questi casuali, portano a comportamenti imprevedibili delle carte di controllo, in quanto la distribuzione della run length non è più la stessa prevista dalla teoria e che invece (approssimativamente) si otterrebbe con parametri stimati con un numero di osservazioni maggiore; l'ARL di fatto è una variabile casuale. **Le stime di parametri non sono i parametri della popolazione**. Lo scenario che risulta più interessante è il successivo in cui i due parametri μ_1 e μ_2 sono ignoti.

2.3 Terzo scenario: nessuno dei parametri è noto

Dato un campione di ampiezza n testiamo la presenza di un unico punto di svolta τ:

$$\begin{cases} H_0 : \tau > n \\ H_1 : \tau \leq n \end{cases}$$

Il sistema di ipotesi è espresso in un contesto di Fase I (quanto detto successivamente sarà ovviamente valido anche per la Fase II). La nostra statistica test

è:

$$\begin{aligned} T_{max,n} &= \log \frac{L_{H_1}(\hat{\mu}_1, \hat{\mu}_2, \hat{\tau})}{L_{H_0}(\hat{\mu}_1)} \\ &= l_{H_1}(\hat{\mu}_1, \hat{\mu}_2, \hat{\tau}) - l_{H_0}(\hat{\mu}_1) \end{aligned}$$

Il processo è in controllo quando siamo sotto H_0, altrimenti sotto H_1. Come si nota quando siamo sotto H_0 l'unico parametro da stimare è μ_1, poichè la log-verosimiglianza in questo caso dipende solo da μ_1. Si ha inoltre che le stime di massima versomiglianza dei parametri (a seconda dell'ipotesi sotto la quale ci troviamo) dato un campione casuale semplice $y = (\ y_1\ y_2\ \dots\ y_n\ )$ sono le seguenti:

$$\hat{\mu}_{1_{H_1}} = \overline{y}_1 \tag{2.1}$$

$$\hat{\mu}_{2_{H_1}} = \overline{y}_2 \tag{2.2}$$

$$\hat{\mu}_{1_{H_0}} = \overline{y} \tag{2.3}$$

Dove:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\ , \quad \overline{y}_1 = \frac{1}{t-1}\sum_{i=1}^{t-1} y_i\ , \quad \overline{y}_2 = \frac{1}{n-t+1}\sum_{i=t}^{n} y_i$$

t non è altro che $\hat{\tau}$ la stima di massima verosimiglianza di τ, o meglio[1]:

$$l_{H_1}(\hat{\mu}_1, \hat{\mu}_2, t) = \max_{2\le j\le n} l_{H_1}(\hat{\mu}_{1j}, \hat{\mu}_{2j}, j) \tag{2.4}$$

Dalle (2.1), (2.2) e (2.3) si ha che:

$$l_{H_1}(\hat{\mu}_1, \hat{\mu}_2, \hat{\tau}) = -(t-1)\log\overline{y}_1 - (n-t+1)\log\overline{y}_2 - n \tag{2.5}$$

$$l_{H_0}(\hat{\mu}_1) = -n\log\overline{y} - n \tag{2.6}$$

Si noti il vincolo lineare:

$$\hat{\mu}_{1_{H_0}} = \frac{(t-1)\,\hat{\mu}_{1_{H_1}} + (n-t+1)\,\hat{\mu}_{2_{H_1}}}{n} \tag{2.7}$$

[1] Si comprende dalla (2.4) anche il motivo della notazione della statistica test $T_{max,n}$, infatti:

$$T_{max,n} = \max_{2\le j\le n} T_{j,n}$$

con $T_{j,n} = l_{H_1}(\hat{\mu}_{1_j}, \hat{\mu}_{2_j}, j) - l_{H_0}(\hat{\mu}_1)$.

Dalle (2.5), (2.6) e (2.7) si ottiene:

$$\begin{aligned} T_{max,n} &= -(t-1)\log\overline{y}_1 - (n-t+1)\log\overline{y}_2 + n\log\overline{y} \\ &= n\log\left(\frac{(t-1)\left(\frac{\overline{y}_1}{\overline{y}_2}\right)^{\frac{n-t+1}{n}} + (n-t+1)\left(\frac{\overline{y}_1}{\overline{y}_2}\right)^{-\frac{t-1}{n}}}{n}\right) \end{aligned} \tag{2.8}$$

Se definiamo $\overline{r} = \overline{y}_1/\overline{y}_2$ allora la statistica $T_{max,n}$ misura in un certo senso la "distanza" di $\overline{r}$ da 1; se infatti il reale rapporto delle medie è 1 (quando siamo sotto H_0) ci aspettiamo che la statistica test $T_{max,n}$ assuma valori vicini allo 0.

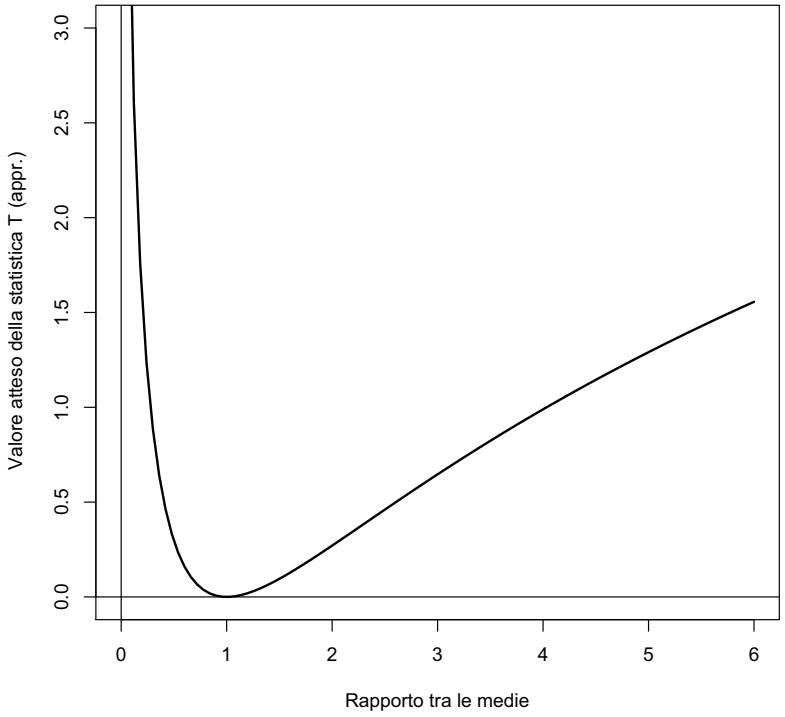

Figura 2.1: Grafico di come approssimativamente la statistica test $T_{max,n}$ misura la "distanza" del rapporto delle medie da 1.

Non siamo nell'ambito dell'inferenza classica: non possiamo difatti invocare la teoria asintotica standard della verosimiglianza. Ricordiamo ancora che applicando lo SPC alla Fase II, ad ogni nuova realizzazione di un tempo di attesa noi effettuiamo sequenzialmente il test per la presenza di un punto di svolta nel processo e se il processo è realmente in cotrollo l'analisi sequenziale non avrà termine. Pertanto la differenza sostanziale che sussite tra carte di

controllo e le usuali verifiche di ipotesi è che in queste ultime una volta ottenuto il valore osservato per la statistica test siamo in grado una volta per tutte di concludere, entro certi limiti (potenza del test), a favore o meno dell'ipotesi nulla (mentre ciò non accade nell'analisi sequenziale, che possiamo interpretare come una serie senza termine di verifiche di ipotesi). Questo fatto ci fa anche comprendere il motivo per il quale è stato rimosso il 2 dalla statistica $T_{max,n}$ dato che per $n \to \infty$ non si ha sotto l'ipotesi nulla la distribuzione asintotica χ^2_1 (un solo g.d.l. poichè faremmo inferenza solo su τ): in Fase I poichè lo stato in controllo viene ipotizzato solamente nel campione a disposizione ed in Fase II essendo l'analisi sequenziale (avremo a che fare con distribuzioni condizionate di $T_{max,n}$ come vedremo avanti).

2.3.1 I limiti di controllo

Sebbene, come chiarito prima, dobbiamo tener conto delle peculiarità della carta di controllo GLR possiamo sfruttare comunque gli strumenti teorici dell'inferenza classica per meglio caratterizzarla. Per poter ottenere i nostri limiti di controllo dobbiamo prima fissare la probabilità dell'errore del primo tipo α.
In Fase I (con il campione di ampiezza fissata ad n) infatti questa è data semplicemente da:

$$P_{H_0}[T_{max,n} > h_{n,\alpha}] = \alpha$$

dove $h_{n,\alpha}$ è il nostro limite con il quale stabiliamo in base al valore osservato della statistica $T_{max,n}$ se l'insieme di dati di cui disponiamo proviene da un processo in controllo e nel qual caso stimiamo il valore di μ_1.
Nella Fase II le cose cambiano. Potremmo iniziare ad effettuare il test dalla terza osservazione di cui disponiamo, ma questo è in realtà molto azzardato poichè con sole tre osservazioni non siamo affatto in grado di poter verificare se i dati provengono da una distribuzione esponenziale e la loro indipendenza. Pare pertanto più opportuno accumulare un certo numero di osservazioni prima di dare il via all'analisi sequenziale dei dati, ed Hawkins propone come numerosità minima di partenza n pari a 10.
In Fase II la probablità dell'errore del primo tipo è data da:

$$P_{H_0}[T_{max,n} > h_{n,\alpha} \mid T_{max,j} \leq h_{j,\alpha} \ , \ j < n] = \alpha$$

In parole α è "la probabilità di lanciare un falso allarme dato che prima non è stato mai lanciato", ossia dobbiamo condizionarci al fatto che in tutti i precedenti test non stati commessi errori del primo tipo. Si nota dunque che la distribuzione della run length se il processo è in controllo è geometrica, infatti come nella Shewart $ARL_0 = 1/\alpha$; equivalentemente possiamo dire che la funzione di rischio della run length è costante e pari ad α [2].
Abbiamo ottenuto via Monte Carlo (si veda la Tabella 2.1) delle stime preliminari dei limiti di controllo. Hawkins nello stimare i limiti di controllo nel caso normale con cambiamenti nella media del processo parte da 16 milioni di

[2] La v.a. geometrica gode della proprietà di mancanza di memoria ed ha una funzione di rischio costante. Infatti:

$$P_{H_0}[RL = n | RL \geq n] = \alpha$$

	α					α			
n	0.05	0.025	0.005	0.001	n	0.05	0.025	0.005	0.001
10	3.816	4.567	6.198	7.846	51	2.873	3.689	5.553	7.405
11	3.339	4.079	5.771	7.461	52	2.871	3.681	5.561	7.381
12	3.13	3.885	5.584	7.301	53	2.856	3.703	5.54	7.325
13	3.025	3.797	5.536	7.321	54	2.867	3.71	5.534	7.334
14	2.962	3.747	5.509	7.212	55	2.881	3.7	5.54	7.32
15	2.926	3.728	5.508	7.217	56	2.885	3.677	5.551	7.338
16	2.901	3.698	5.494	7.235	57	2.855	3.669	5.576	7.38
17	2.896	3.69	5.505	7.24	58	2.881	3.691	5.56	7.353
18	2.883	3.694	5.485	7.208	59	2.883	3.669	5.569	7.378
19	2.875	3.664	5.503	7.186	60	2.883	3.7	5.555	7.327
20	2.864	3.668	5.482	7.246	61	2.878	3.683	5.538	7.342
21	2.863	3.669	5.502	7.233	62	2.863	3.7	5.56	7.339
22	2.858	3.672	5.517	7.232	63	2.862	3.686	5.59	7.313
23	2.861	3.672	5.505	7.267	64	2.895	3.686	5.555	7.366
24	2.865	3.665	5.506	7.26	65	2.868	3.703	5.572	7.4
25	2.862	3.672	5.524	7.269	66	2.891	3.669	5.575	7.351
26	2.873	3.656	5.528	7.248	67	2.917	3.693	5.561	7.383
27	2.856	3.672	5.5	7.236	68	2.871	3.706	5.568	7.374
28	2.857	3.666	5.501	7.246	69	2.906	3.665	5.564	7.398
29	2.863	3.655	5.505	7.274	70	2.882	3.685	5.538	7.345
30	2.857	3.675	5.521	7.351	71	2.918	3.698	5.552	7.405
31	2.85	3.678	5.519	7.285	72	2.874	3.687	5.571	7.386
32	2.844	3.655	5.522	7.215	73	2.88	3.682	5.553	7.397
33	2.859	3.677	5.543	7.276	74	2.859	3.686	5.568	7.421
34	2.867	3.672	5.528	7.302	75	2.869	3.709	5.58	7.376
35	2.86	3.682	5.552	7.281	76	2.902	3.705	5.57	7.409
36	2.854	3.676	5.52	7.292	77	2.892	3.708	5.591	7.44
37	2.874	3.671	5.563	7.259	78	2.859	3.716	5.584	7.391
38	2.864	3.668	5.541	7.307	79	2.83	3.709	5.603	7.366
39	2.858	3.691	5.536	7.299	80	2.854	3.726	5.596	7.374
40	2.863	3.673	5.552	7.323	81	2.901	3.706	5.565	7.347
41	2.855	3.667	5.537	7.359	82	2.884	3.693	5.578	7.361
42	2.856	3.693	5.542	7.296	83	2.886	3.7	5.574	7.382
43	2.861	3.683	5.532	7.294	84	2.834	3.708	5.577	7.333
44	2.879	3.695	5.555	7.287	85	2.833	3.691	5.559	7.338
45	2.856	3.67	5.538	7.289	86	2.927	3.668	5.578	7.329
46	2.87	3.708	5.531	7.338	87	2.855	3.689	5.558	7.356
47	2.871	3.705	5.546	7.377	88	2.832	3.671	5.579	7.349
48	2.885	3.689	5.511	7.34	89	2.889	3.71	5.574	7.332
49	2.867	3.68	5.545	7.308	90	2.921	3.713	5.554	7.4
50	2.883	3.669	5.569	7.378					

Tabella 2.1: Stime via simulazione dei limiti di controllo dato un livello di confidenza α e numerosità n.

n	$h_{n,0.025}$	$e_{perc.}$	n	$h_{n,0.025}$	$e_{perc.}$	n	$h_{n,0.025}$	$e_{perc.}$	n	$h_{n,0.025}$	$e_{perc.}$
10	4.553	0.092	60	3.69	0.124	110	3.696	0.218	160	3.708	0.441
11	4.082	0.069	61	3.69	0.117	111	3.698	0.238	161	3.707	0.453
12	3.888	0.062	62	3.691	0.135	112	3.698	0.226	162	3.702	0.403
13	3.791	0.059	63	3.69	0.125	113	3.696	0.229	163	3.7	0.441
14	3.74	0.062	64	3.691	0.138	114	3.694	0.235	164	3.701	0.471
15	3.711	0.065	65	3.691	0.138	115	3.691	0.238	165	3.703	0.466
16	3.695	0.065	66	3.69	0.132	116	3.694	0.255	166	3.705	0.358
17	3.684	0.068	67	3.692	0.133	117	3.698	0.262	167	3.698	0.459
18	3.678	0.07	68	3.693	0.126	118	3.696	0.257	168	3.701	0.482
19	3.674	0.068	69	3.692	0.143	119	3.695	0.262	169	3.689	0.479
20	3.672	0.065	70	3.691	0.148	120	3.698	0.256	170	3.693	0.464
21	3.671	0.068	71	3.692	0.13	121	3.694	0.269	171	3.704	0.48
22	3.671	0.068	72	3.691	0.144	122	3.695	0.244	172	3.695	0.468
23	3.671	0.064	73	3.692	0.165	123	3.696	0.276	173	3.701	0.523
24	3.671	0.072	74	3.691	0.153	124	3.697	0.266	174	3.697	0.496
25	3.671	0.075	75	3.693	0.16	125	3.694	0.295	175	3.691	0.449
26	3.672	0.077	76	3.693	0.151	126	3.701	0.281	176	3.692	0.533
27	3.673	0.082	77	3.692	0.177	127	3.693	0.286	177	3.699	0.523
28	3.673	0.08	78	3.692	0.144	128	3.698	0.283	178	3.688	0.545
29	3.673	0.09	79	3.694	0.148	129	3.696	0.277	179	3.7	0.552
30	3.675	0.073	80	3.691	0.144	130	3.696	0.306	180	3.701	0.536
31	3.676	0.077	81	3.696	0.176	131	3.698	0.295	181	3.696	0.496
32	3.676	0.092	82	3.69	0.156	132	3.7	0.35	182	3.695	0.553
33	3.677	0.089	83	3.695	0.155	133	3.697	0.315	183	3.692	0.548
34	3.678	0.083	84	3.693	0.158	134	3.698	0.297	184	3.699	0.554
35	3.68	0.089	85	3.692	0.15	135	3.701	0.354	185	3.711	0.585
36	3.679	0.093	86	3.692	0.177	136	3.7	0.349	186	3.713	0.678
37	3.68	0.09	87	3.694	0.183	137	3.698	0.362	187	3.707	0.597
38	3.68	0.095	88	3.695	0.169	138	3.695	0.332	188	3.7	0.55
39	3.682	0.097	89	3.695	0.169	139	3.698	0.339	189	3.702	0.594
40	3.682	0.097	90	3.698	0.163	140	3.695	0.359	190	3.7	0.59
41	3.682	0.092	91	3.694	0.201	141	3.693	0.402	191	3.701	0.647
42	3.683	0.103	92	3.694	0.187	142	3.695	0.401	192	3.709	0.593
43	3.683	0.1	93	3.696	0.185	143	3.701	0.345	193	3.707	0.685
44	3.683	0.109	94	3.698	0.18	144	3.7	0.354	194	3.709	0.661
45	3.684	0.093	95	3.698	0.207	145	3.698	0.338	195	3.716	0.669
46	3.687	0.087	96	3.697	0.182	146	3.696	0.383	196	3.708	0.663
47	3.686	0.097	97	3.694	0.19	147	3.703	0.375	197	3.697	0.669
48	3.686	0.1	98	3.697	0.209	148	3.702	0.331	198	3.7	0.655
49	3.685	0.099	99	3.695	0.219	149	3.697	0.358	199	3.702	0.73
50	3.686	0.102	100	3.696	0.193	150	3.703	0.34	200	3.698	0.763
51	3.686	0.114	101	3.691	0.189	151	3.696	0.356			
52	3.688	0.107	102	3.696	0.203	152	3.701	0.399			
53	3.686	0.121	103	3.695	0.194	153	3.706	0.368			
54	3.687	0.115	104	3.698	0.207	154	3.695	0.373			
55	3.687	0.111	105	3.697	0.215	155	3.7	0.427			
56	3.69	0.124	106	3.694	0.229	156	3.7	0.436			
57	3.69	0.117	107	3.697	0.245	157	3.705	0.445			
58	3.69	0.12	108	3.699	0.205	158	3.693	0.414			
59	3.69	0.128	109	3.698	0.215	159	3.699	0.478			

Tabella 2.2: Per ogni n viene riportata la stima del limite di controllo con livello di confidenza $\alpha = 0.025$, ottenuto dalla media campionaria di 100 stime, e relativo errore percentuale.

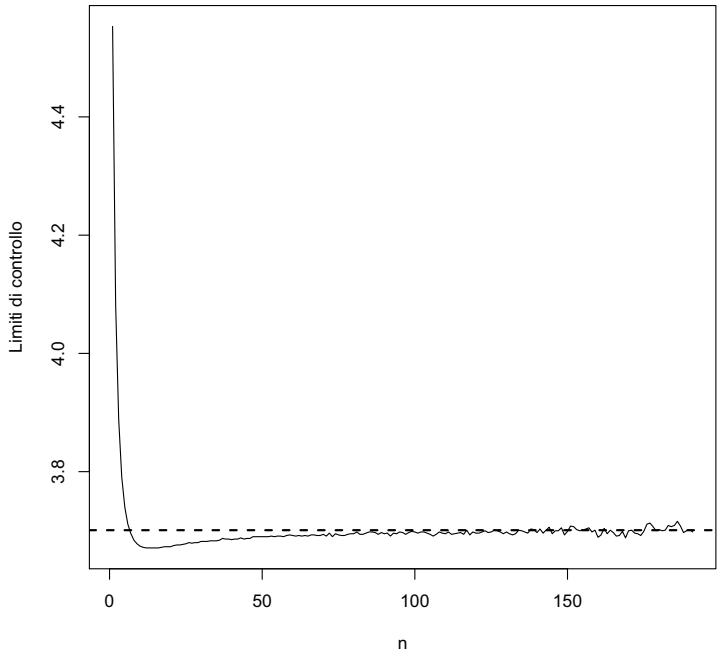

Figura 2.2: Grafico dei limiti di controllo riportati in Tabella 2.2 verso n.

osservazioni, le nostre stime invece sono ottenute partendo da 500 mila osservazioni. Ci siamo così concentrati su un particolare livello di confidenza, ossia $\alpha = 0.025$, e partendo sempre da 500 mila osservazione abbiamo ripetuto l'esecuzione dell'algoritmo per 100 volte (si veda la Tabella 2.2) spingendoci fino ad $n = 200$. Abbiamo avuto così a disposizione per ogni n, 100 stime dei limiti di controllo. Abbiamo poi fatto la media campionaria in ogni gruppo di 100 valori (la media campionaria è uno stimatore corretto). Si fa notare inoltre che nella simulazione al progredire di n disponiamo sempre di meno osservazioni, poichè dobbiamo condizionarci ai campioni in cui non ci sono stati falsi allarmi, quindi la variabilità delle stime sarà sempre più grande all'aumentare di n (ad incidere su tale variabilità entra in gioco anche α poichè più è grande più la variabilità sarà maggiore a parità di n); per $n = 200$ abbiamo infatti circa $500000 \times 0.975^{200} \approx 3160$ osservazioni per stimare il limite di controllo, di fatto un numero non molto alto. Inoltre nella Tabella 2.2 se avessimo voluto ridurre di un ulteriore ordine di grandezza l'errore percentuale avremmo dovuto ripetere per 10000 volte l'esecuzione dell'algoritmo, dato che la deviazione standard è $\sigma_{B,\alpha}/\sqrt{m}$ (deriva dal fatto che facciamo la media campionaria di m valori; $\sigma_{B,\alpha}$ è la deviazione standard dello stimatore del quantile-α calcolato su B osservazioni), per questo motivo ci siamo fermati ad $m = 100$.
Come si vede dalla Figura 2.2 i limiti di controllo al divergere di n convergono verso un valore che sembra essere non molto più grande di 3.7 e che può essere utilizzato come valore approssimato dei limiti a partire circa da $n = 100$.

Dunque i limiti di controllo al divergere di n si stabilizzano. A conferma di quanto detto precendentemente vediamo che l'errore percentuale aumenta con n, ciò si riflette in Figura 2.2 con il maggior "rumore" che si ha verso n grandi.

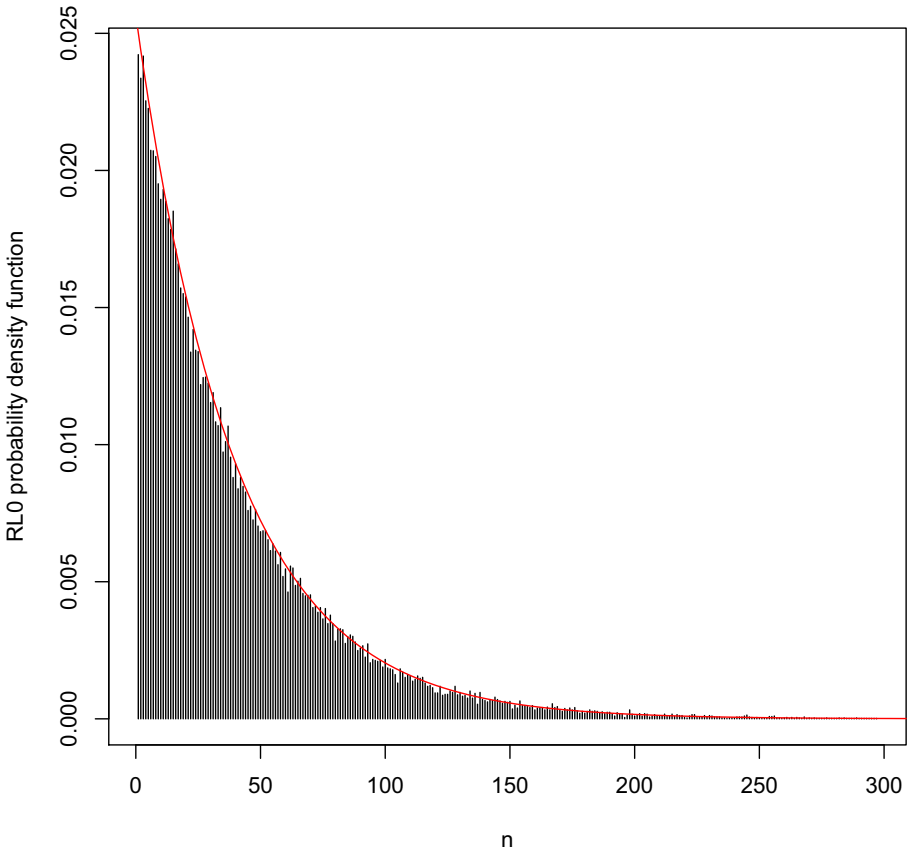

Figura 2.3: Le linee verticali nere rappresentano la funzione di probabilità stimata della run length in controllo; la linea continua in rosso rappresenta funzione di probabilità di una v.a. geometrica con $p = 0.025$ (la continuità della linea rossa non ha senso ma rende visibilmente più facile l'interpretazione).

Ci chiediamo ora quanto le stime dei limiti di controllo siano buone, non solo in termini di errore percentuale, e se l'apporossimazione dei limiti di controllo al valore 3.7 per $n \geq 100$ sia sensata oppure no. Per rispondere possiamo stimare la funzione di probabilità della run length in controllo e vedere se è geometrica oppure se ci sono chiari discostamenti da questa assunzione. Abbiamo operato in tal modo a partire da $n = 10$ con 200 mila osservazioni e come si vede anche dalla Figura 2.3 (anche ripetendo più volte la simulazione) la distribuzione della run length con quei limiti di controllo stimati risulta essere approssimativamente una v.a. geometrica e notiamo inoltre come per $n \geq 100$ approssimando i limiti

a 3.7 il comportamento della run length è chiaramente identico a quello di una v.a. geometrica.

2.3.2 Performance della carta di controllo

Come abbiamo visto la run length in controllo segue una distribuzione geometrica. Fuori controllo invece la distribuzione della run length non è più geometrica, ossia la funzione di rischio diventa più genericamente:

$$1 - \beta_{n,\tau,r} = \frac{P_{H_1}[RL = n]}{P_{H_1}[RL \geq n]} \ , \ n \geq \tau$$

la probabilità di commettere un errore del II tipo dipende dunque oltre che da n, anche dal punto di svolta τ e dal vero rapporto tra le medie $r = \mu_1/\mu_2$. Supponiamo ora che dopo il punto di svolta τ ci sia un cambiamento nella media da μ_1 a μ_2 (il cui rapporto come definito prima è r) e calcoliamo il parametro di non centralità:

$$\begin{aligned} E_{H_1}[T_{max,n} \mid \tau] &= E_{H_1}[-(\tau-1)\log\overline{Y}_1 - (n-\tau+1)\log\overline{Y}_2 + n\log\overline{Y}] \\ &= -E_{H_1}[(\tau-1)\log\overline{Y}_1] - E_{H_1}[(n-\tau+1)\log\overline{Y}_2] \ + \\ &\quad + E_{H_1}[n\log\overline{Y}] \end{aligned}$$

Poichè non è possibile esprimere in forma chiusa il calcolo dei tre valori attesi, si effettuano tre sviluppi di Taylor ognuno centrato rispettivamente in μ_1, μ_2 e l'ultimo in $((\tau-1)\mu_1 + (n-\tau+1)\mu_2)/n$ (che non sono altro che i valori attesi rispettivamente di $\overline{Y}_1$, $\overline{Y}_2$ e di $\overline{Y}$). Cosi abbiamo che il parametro di non centralità risulta essere approssimativamente:

$$E_{H_1}[T_{max,n} \mid \tau] \doteq -(\tau-1)\log r + n\log\left(1 \ + \ \frac{(\tau-1)(r-1)}{n}\right) \qquad (2.9)$$

Mentre sotto H_0 la statistica test $T_{max,n}$ dipende solo da n, sotto H_1 dipende oltre che da n anche da τ ed r.
Il parametro di non centralità, al divergere di n, raggiunge un limite che approssimativamente risulta essere:

$$(\tau-1)(r - \log r - 1) \qquad (2.10)$$

Questo suggerisce che la funzione di rischio della run length fuori controllo incrementa dopo che il cambiamento è occorso, per poi stabilizzarsi su un qualche livello. Come si osserva dalla Figura 2.4 la funzione di rischio (stimata a partire da 200 mila osservazioni con $\tau = 10$ e $r = 2$) raggiunge un massimo per poi decrescere e stabilizzarsi: intuiamo perciò che le code della distribuzione della run length fuori controllo non si discostano molto da quelle di una distribuzione geometrica. L'ARL rappresenta un sensato criterio di confronto con altre carte di controllo quali ad esempio la CUSUM (se la funzione di rischio fosse costante la distribuzione della run length sarebbe completamente carat-

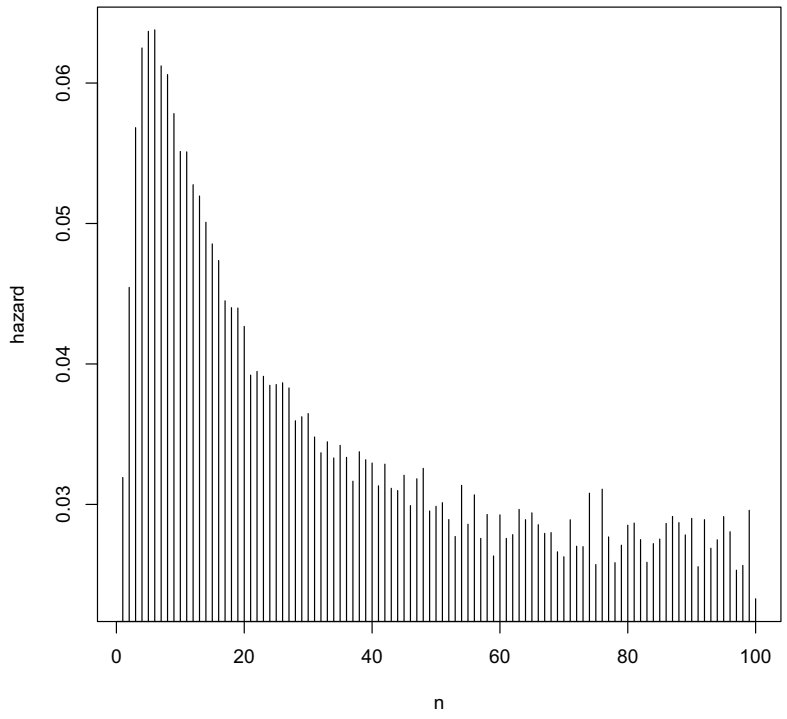

Figura 2.4: La funzione di rischio della run length fuori controllo (stimata).

terizzata dall'ARL). Osserviamo inoltre che esiste una "finestra di opportunità" negli istanti successivi al cambiamento in cui è più probabile che lo stesso venga rilevato: o ci accorgiamo quasi subito del cambiamento o è più difficile (meno probabile) accorgersene dopo.
In Tabella 2.3 è riportato il profilo dell'ARL: una fatto molto evidente è come l'ARL diminuisce all'aumentare di τ, ossia più è grande la storia del processo in controllo e più rapidamente saremo in grado di rilevare, una volta che il processo va fuori controllo, il cambiamento; infatti all'aumentare di τ aumenta il parametro di non centralità.

	$\frac{\mu_2}{\mu_1}$										
τ	0.25	0.5	0.75	1	1.5	2	2.5	3	4	7	13
10	7.7	23.1	35.1	40.1	37.6	29.7	21.9	15.7	8.5	3.3	2.0
25	5.2	15.6	31.6	39.9	32.2	18.3	10.5	7.0	4.3	2.4	1.6
50	4.8	12.6	28.9	39.9	27.5	13.3	8.0	5.7	3.8	2.2	1.6
100	4.7	11.3	26.3	40.4	23.8	11.6	7.1	5.3	3.6	2.2	1.6
200	4.6	10.7	25.1	38.2	21.2	10.4	6.7	5.0	3.5	2.2	1.5

Tabella 2.3: ARL in controllo e fuori controllo quando la probabilità di commettere un errore del I tipo è fissata ad $\alpha = 0.025$.

Nella colonna in cui il rapporto tra le medie è pari ad 1, osserviamo l'ARL in controllo che sappiamo essere pari ad $1/0.025 = 40$, poichè ricordiamo che la distribuzione della run length in controllo è geometrica.

2.4 Esempio: disastri in una miniera di carbone

Analizzaremo un dataset disponibile nella libreria *boot* di R sotto il nome di *coal*, in cui sono riportate le date di 191 esplosioni in una miniera di carbone in ognuna della quali ci sono state 10 o più morti sul posto, dal 15 Marzo 1851 al 22 Marzo 1962:

	date
1	1851.203
2	1851.632
3	1851.969
4	1851.975
5	1852.314
6	1852.347
7	1852.358
8	1852.385
9	1852.977
10	1853.196
.	.
.	.
.	.
185	1947.619
186	1947.638
187	1947.687
188	1951.405
189	1957.883
190	1960.489
191	1962.220

Facciamo un grafico del numero totale di disastri occorso in ogni anno dal 1851 al 1962: si può osservare che dall'anno 1900 circa in poi il numero di disastri annui mediamente è diminuito e si nota anche una minor variabilità (come è

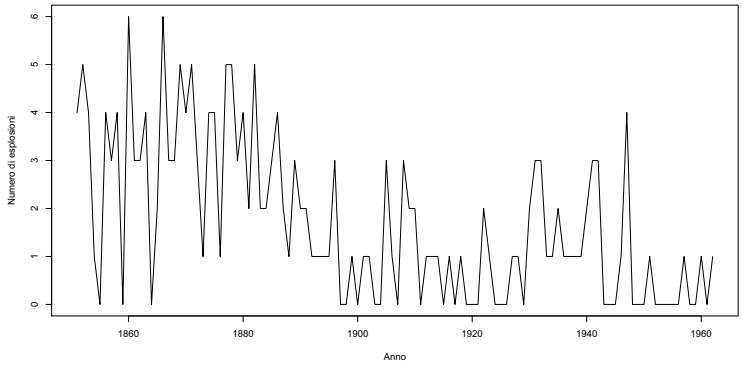

Figura 2.5: Andamento nel tempo del numero di disastri annui.

logico aspettarsi con dati di Poisson). Per poter applicare le carte di controllo presentate fin qui dobbiamo ricavare dal dataset i tempi che intercorrono tra due disastri consecutivi che in totale saranno dunque 190:

	TBD
1	0.4298
2	0.3368
3	0.0055
4	0.3395
5	0.0329
6	0.0110
7	0.0274
8	0.5914
9	0.2190
10	0.0329
.	.
.	.
.	.
185	0.0192
186	0.0493
187	3.7180
188	6.4778
189	2.6064
190	1.7303

Per verificare l'assunto distributivo e l'indipendenza supponiamo di disporre

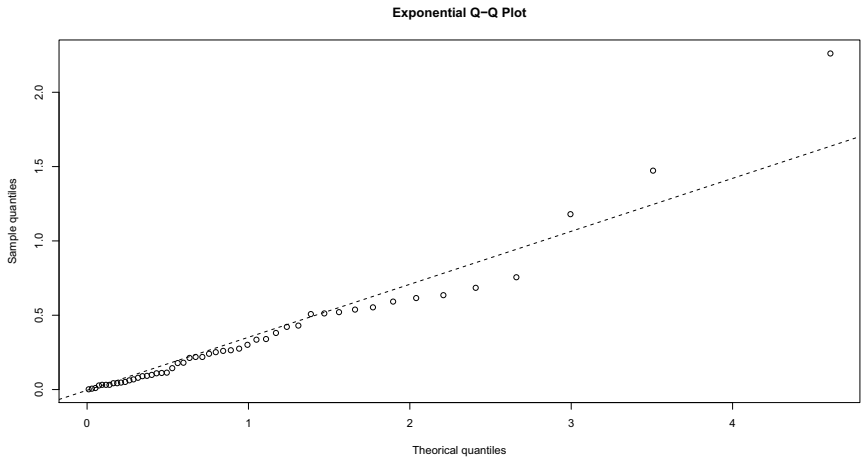

Figura 2.6: Q-Q plot per le prime 50 osservazioni.

delle prime 50 osservazioni, sulle quali ci baseremo anche per una stima preliminare di μ_1 (come si vedrà in seguito in queste prime 50 osservazioni il processo risulterà essere in controllo). Dal Q-Q plot osserviamo che non sembrano esserci scostamenti allarmanti dalla retta che interpola il primo decile ed il terzo quartile, se non in parte nella zona centrale e finale. Inoltre con il test di esponenzialità di Shapiro-Wilk, fissato $\alpha = 0.05$, si ha che (il test è bilaterale):

$$q_{50,0.025} = 0.0122 \ < \ t_{S.W.oss.} = 0.0136 \ < \ q_{50,0.975} = 0.0340$$

i dati sono conformi rispetto all'ipotesi nulla di esponenzialità, anche se ovviamente non possiamo dire che provengono da una v.a. esponenziale.

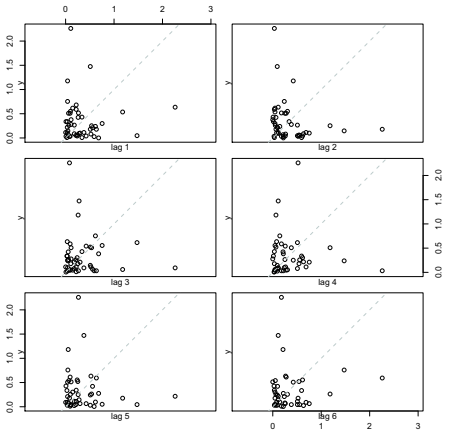

Figura 2.7: Lag-plot dei dati fino a ritardo 6.

Per quanto riguarda la verifica dell'assunto di indipendeza i dati risultano essere

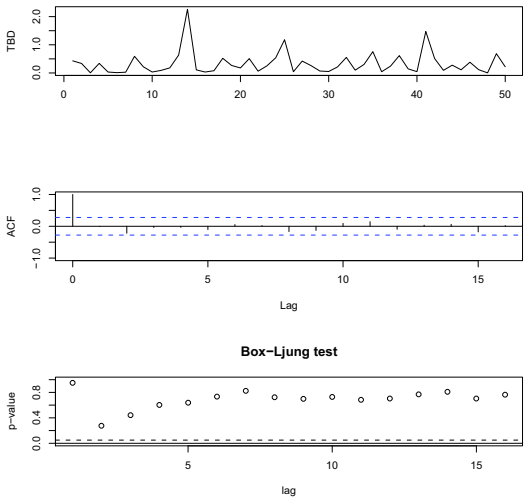

Figura 2.8: Dall'alto verso il basso: rappresentazione grafica dei dati; correlogramma; p-value del test di Ljung-Box a vari ritardi.

incorrelati come si nota dal correlogramma e dai p-value del test di Box-Ljung (infatti le autocerrelazioni sono prossime allo 0 ed i p-value sono tutti molto grandi per opportuni ritardi). Da segnalare è la presenza di un'apparente stagionalità che si evince nell'andamento dei dati nel tempo (primo pannello di Figura 2.7) ed anche nei "cerchi" dei lag-plot (Figura 2.8). Volendo verificare l'eventuale dipendenza dei dati da mese a mese e da giorno a giorno, identifichiamo a partire dalla parte frazionaria delle date i giorni ed i mesi. Tale approccio è anche seguito da Lucas (1985) nell'analisi di un famoso dataset delle morti sul lavoro. Abbiamo dunque le seguenti tabelle di frequenza:

Mese	N. Esplosioni
Gennaio	14
Febbraio	20
Marzo	20
Aprile	13
Maggio	14
Giugno	10
Luglio	17
Agosto	16
Settembre	10
Ottobre	16
Novembre	17
Dicembre	24
$\chi^2_{oss.} \doteq 11.7435 - \alpha_{oss.} \doteq 0.3832$	

Giorno	N. Esplosioni
Lunedì	10
Martedì	29
Mercoledì	33
Giovedì	33
Venerdì	37
Sabato	31
Domenica	18
$\chi^2_{oss.} \doteq 20.5759 - \alpha_{oss} \doteq 0.002186$	

Mediante tali tabelle verifichiamo il seguente sistema di ipotesi:

$$\begin{cases} H_0 : \textit{Indipendenza stocastica da mese a mese} \\ H_1 : \textit{Esiste una qualche relazione da mese a mese} \end{cases}$$

la statistica-test con relativa distribuzione nulla ottenuta asintoticamente sono:

$$\sum_{i=1}^{12} \frac{\left(f_i - \hat{f}\right)^2}{\hat{f}} \underset{H_0}{\dot{\sim}} \chi^2_{11}$$

dove f_i sono le frequenze osservate ed $\hat{f}$ le frequenze attese. Nel caso dei giorni si hanno un analogo sistema di ipotesi e statistica test (ovviamente avremo 6 gradi di libertà invece che 11). Nel caso dei mesi la frequenza attesa (sotto l'ipotesi di indipendenza) risulta essere $191/12 \approx 15.92$ mentre nel caso dei giorni risulta essere $191/7 \approx 27.29$. I valori dei p-value ci suggeriscono che mentre sembrerebbe esserci indipendenza da mese a mese ciò non è ancora valido da giorno a giorno. Infatti le esplosioni si concentrano maggiormente dentro la settimana mentre la domenica ed il lunedì sono meno frequenti. Questo fatto potrebbe essere dovuto al particolare tipo di orario lavorativo osservato durante la settimana in miniera, tenendo conto anche dei cambiamenti che possono essere avvenuti dal 1851 al 1922 (sarebbe infatti opportuno addentrarci nella storia della nostra mineria di carbone e capire come si sono evolute le misure di sicurezza in modo da chiarire la diminuzione del numero medio di esplosioni annue). In linea del tutto illustrativa trascuriamo questa dipendenza da giorno a giorno anche perchè come si vedrà le nostre carte di controllo riusciranno comunque nel caso in esame ad assolvere il loro compito di segnalazione del cambiamento.

La stima di cui disponiamo per la media in controllo è $\hat{\mu}_1 = 0.333$, indispensabile nel disegno delle carte Shewart e CUSUM. Questa seconda carta sarà disegnata in maniera tale da essere ottimale nella segnalazione di aumenti pari a 1.5 volte la media in controllo e diminuzioni pari a 0.6 volte la media in controllo. Uno dei vantaggi che ci fornisce la carta GLR è che non si ha bisogno di alcuna stima preliminare di μ_1 per la sua applicazione. L'ARL in controllo è fissato a 200 (per tutte e tre le carte).

Osservando i grafici delle tre carte di controllo possiamo affermare che nel caso in esame si comportano tutte e tre bene. Avevamo osservato nell'anda-

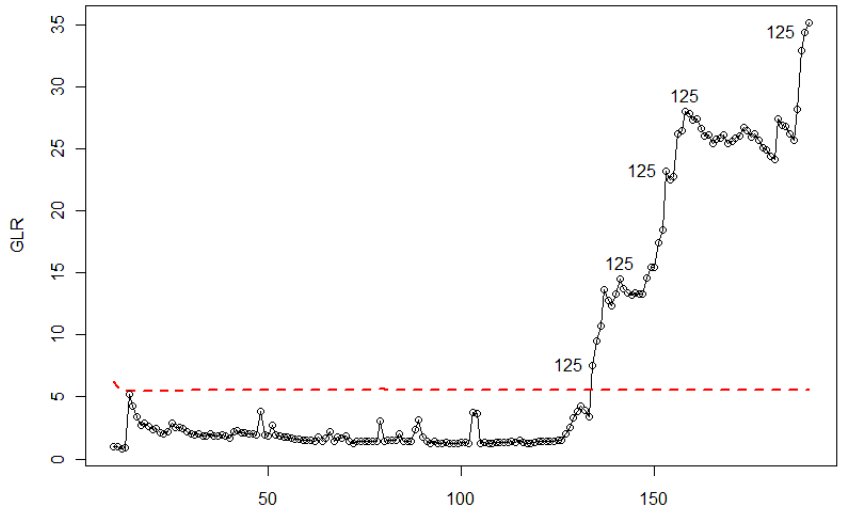

Figura 2.9: La carta di controllo GLR (si monitora il processo a partire dalla decima osservazione); i numeri sopra i punti sono le stime di massima verosigmiglianza di τ nei vari istanti successivi alla segnalazione del fuori controllo.

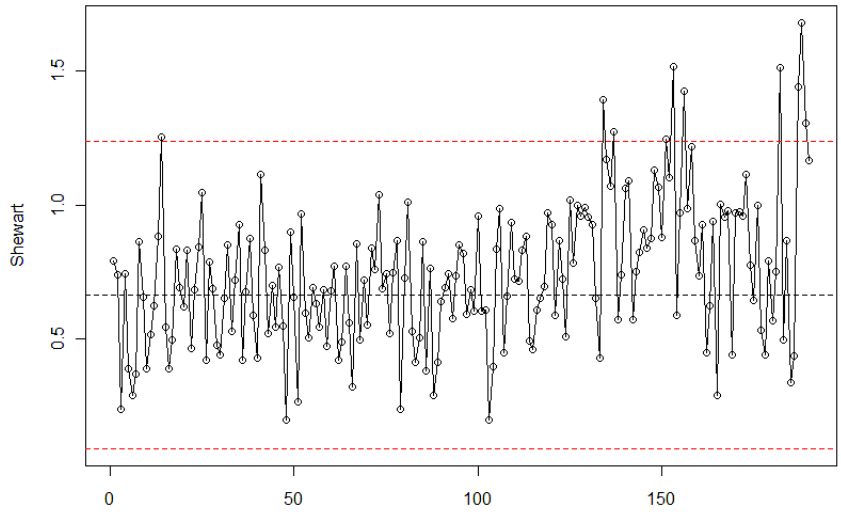

Figura 2.10: La carta di controllo Shewart (con limiti esatti).

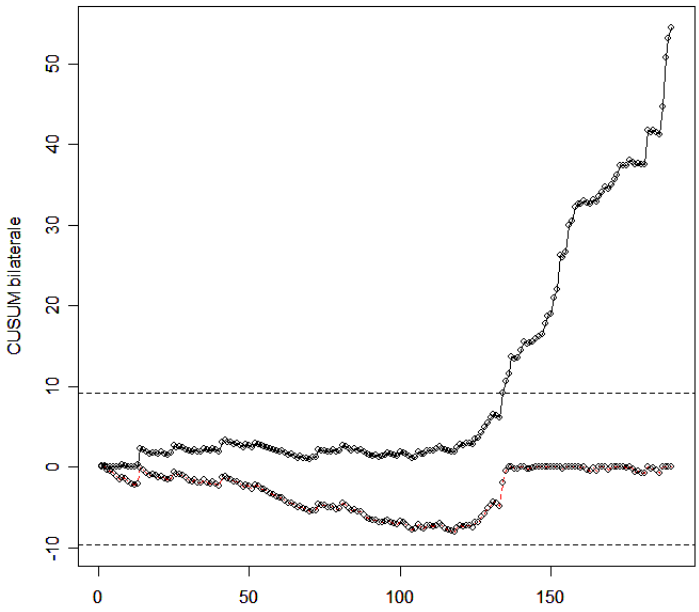

Figura 2.11: La carta di controllo Cusum bilaterale.

mento del numero annuo di esplosioni che c'è una diminuzione media intorno all'inizio del XX secolo. Questo comporta un aumento nella media del tempo di attesa medio tra un esplosione e un'altra come infatti si rileva dal grafico delle carte di controllo Shewart e CUSUM nella deriva verso l'alto (non è osservabile questo fatto nell'andamento della carta GLR poichè anche ci fosse stata una diminuzione nella media la deriva sarebbe stata comunque verso l'alto, ma ciò è ovvio perchè tale carta è disegnata per rilevare cambiamenti nella media e qualsiasi sia la direzione del cambiamento l'ipotesi nulla viene rifiutata sempre per valori alti della statistica test). Le tre carte segnalano il cambiamento sostanzialmente con la stessa velocità.
La stima di massima verosimiglianza di τ è 125 mentre per μ_1 e μ_2 risultano essere rispettivamente 0.314 e 1.091: la media in controllo di fatto è triplicata. Dal grafico dell'ARL per la carta Shewart (Figura 1.5, linea trattegiata rossa) siamo nella zona in cui il comportamento della carta è molto buono (infatti $r \doteq 1/3$ e questo spiega la celerità della carta nel segnalare l'aumento della media; da notare il falso allarme che viene lanciato alla quindicesima osservazione. Il cambiamento che osserviamo è di fatto un cambiamento grande ed è difficile riuscire a rilevare da questo unico esempio quale delle tre carte ha una prestazione migliore.

Conclusioni

Abbiamo visto che tra le carte di controllo proposte per il monitoraggio del tempo intercorrente tra due eventi, di cui siamo interessati a rilevare un cambiamento nella media e nel caso in cui la distribuzione sottostante sia esponenziale, la carta di controllo basata sul modello di change-point è un valido strumento e di facile applicazione in quanto non necessità di un iniziale studio di Fase I. Ciò è notevole in particolar modo in contesto industriale (e non solo) dove è sempre più frequente l'avvio di nuovi processi produttivi per i quali non si dispone di un previo studio in controllo. Le altre carte di controllo necessitano invece di uno studio di Fase I per ottenere delle stime dei parametri di interesse e proprio perchè sono delle stime queste possono a volta anche rendere non del tutto efficiente il monitoraggio futuro.
La carta di controllo GLR è in grado di competere con le altre già note (Shewart, CUSUM, EWMA, etc...) e ne rappresenta una più che valida alternativa.

Appendice A

Codice R

Viene qui riportato il codice in linguaggio R sviluppato per ogni parte della tesi.

Q-Q Plot per dati esponenziali

La funzione è stata creata per poter disporre dell'analoga funzione *qqnorm* disponibile in R ma per dati esponenziali.

```
qqexp <- function(dati) {
   qqplot(qexp(ppoints(dati)),sort(dati),
          ylab = "Sample quantiles",xlab = "Theorical quantiles",
          main = " Exponential Q-Q Plot ")
   q1 <- quantile(dati,0.1)
   q3 <- quantile(dati,0.75)
   x1 <- qexp(0.1)
   x3 <- qexp(0.75)
   m <- (q3-q1)/(x3-x1); q <- q1-m*x1
   abline(q,m,lty = "dashed")
}
```

La retta riportata nel grafico interpola il primo decile ed il terzo quartile in modo da seguire meglio l'asimmetria dei dati.

Quantili della statistica $T_{max,n}$

L'algoritmo mediante il quale sono stati stimati i quantili della statistica $T_{max,n}$ riportati nelle due tabelle del captiolo 2.

```
### La funzione che restituisce il massimo del
### log-rapporto di verosimiglianza per ogni colonna (in totale B)
### della matrice somme.cum (matrice delle somme cumulate)

max.glr <- function() {
      # Il vettore degli istanti di cambiamento
      t <- 2:n
      # Calcola la statistica T(max,n)
      calcolo.glr <- function(vsomme) {
           y.bar1 <- vsomme[t-1]/(t-1)
           y.bar2 <- (vsomme[n]-vsomme[t-1])/(n-t+1)
           max(-(t-1)*log(y.bar1)-(n-t+1)*log(y.bar2))+n*log(vsomme[n]/n)
      }
      sapply(1:(dim(somme.cum)[2]),function(j) calcolo.glr(somme.cum[,j]))
}

start <- 10
end <- 200
B <- 500000
dati <- matrix(rexp(start*B), ncol = B, nrow = start)
somme.cum <- apply(dati, 2,cumsum)
rm(dati)
alpha <- c(0.975)
n <- dim(somme.cum)[1]
quantili <- double(n)

### Primo passo
glr.oss <- max.glr()
quantili[1] <- quantile(glr.oss,alpha)
somme.cum <- somme.cum[,which(glr.oss <= quantili[1])]

### Dal secondo passo all'ultimo
for(i in 2:(end-start+1)) {
```

```
    nuova.sim <- rexp(dim(somme.cum)[2])
    somme.cum <- rbind(somme.cum,somme.cum[n,]+nuova.sim)
    n <- n + 1
    glr.oss <- max.glr()
    quantili[i] <- quantile(glr.oss,alpha)
    somme.cum <- somme.cum[,which(glr.oss <= quantili[i])]
    cat("\n Concluso  passo numero ",i," per n = ",n,"\n")
}
```

Calcolo dei giorni della settimana del data-set *coal*

Viene riportato il codice necessario per ricavare i giorni della settimana in cui ci sono state le esplosioni.

```
library(boot)
data(coal)
x <- coal[,1]
diff.anno <- function(x, rif = 1851) {
    anno <- trunc(x)
    anno.res <- x%%1

    ### Secondi in un anno non bisestile
    anno.sec <- 365*24*60*60

    ### Secondi in un anno bisestile
    anno.sec.bis <- 366*24*60*60

    if(anno == rif) {
       anni.prec <- 0
    } else {
       v.anni <- seq(rif,anno-1,1)
       anni.prec <- sum(((v.anni%%4 == 0 & v.anni%%100 != 0) |
                         (v.anni%%100 == 0 & v.anni%%400 == 0))
                         *anno.sec.bis +
                        ((v.anni%%4 != 0) | (v.anni%%100 == 0 &
```

```
                        v.anni%%400 != 0))*anno.sec)
    }
    anno.res.sec <- sum(((anno%%4 == 0 & anno%%100 != 0) |
                        (anno%%100 == 0 & anno%%400 == 0))
                        *anno.sec.bis*anno.res +
                        ((anno%%4 != 0) | (anno%%100 == 0 &
                        anno%%400 != 0))*anno.sec*anno.res)

                        anni.prec+anno.res.sec
}

z <- sapply(x,diff.anno)

date <- strptime("1851-01-01 GMT", "%Y-%m-%d",tz = "GMT") + z

### Tabella di frequenza del numero di
### esplosioni nei giorni della settimana
table(weekdays(date))
```

Una verifica della correttezza dell'output è che il primo giorno di *date* è il 15 Marzo 1851 e l'ultimo è il 22 Marzo 1962.

Bibliografia

[1] Gan, F. F., *Designs of Optimal Exponential CUSUM Control Charts*, Journal of Quality Technology, 1998

[2] Gan, F. F., *Designs of One- and Two-Sided Exponential EWMA Charts*, Journal of Quality Technology, 1998

[3] Hawkins, D. M., Qui, P., e Kang, C. W., *The Changepoint Model for Statistical Process Control*, Journal of Quality Technology, 2003

[4] Hawkins, D. M. e Zamba, K. D., *Statistical Process Control for Shifts in Mean or Variance Using a Changepoint Formulation*, Technometrics, 2005

[5] Lucas, J. M., *Counted Data CUSUM's*, Technometrics, 1985

[6] Montgomery, D. C., *Controllo statistico della qualità*, McGraw-Hill, Seconda Edizione

[7] Nelson, L. S., *A Control Chart for Parts-Per-Milion Nonconforming Items*, Journal of Quality Technology, 1994

[8] Nelson, L. S., *Constructing Normal Probability Paper*, Journal of Quality Technology, 1976