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

Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

STUDIO E APPLICAZIONE DELLA CARTA DI CONTROLLO NON PARAMETRICA DI FASE II BASATA SU UNA TRASFORMAZIONE DELLO SCORE DI VEROSIMIGLIANZA PER MONITORARE CONGIUNTAMENTE POSIZIONE E SCALA

Relatore Prof. Guido Masarotto
Dipartimento di Scienze Statistiche

Laureando: Federico Passuello
Matricola N. 1220643

Anno Accademico 2023/2024

INDICE

INTRODUZIONE

Il controllo statistico della qualità si occupa di monitorare e valutare la qualità di processi produttivi o di servizi forniti. Per fare ciò ha bisogno di certi strumenti statistici, al fine di garantire che il processo osservato rispetti certi standard di qualità. Oltre a monitorare il processo, cercando di segnalare il prima possibile un allarme se qualcosa di imprevisto è successo, il controllo statistico della qualità, (CSQ) si occupa di fornire specifiche indicazioni correttive e preventive per migliorare continuamente la qualità complessiva del processo. Per raggiungere questi obiettivi, lo strumento base del CSQ è la carta di controllo (CC): uno strumento grafico per valutare se al passare del tempo un processo è rimasto “in controllo” (IC, ossia funziona in modo prevedibile e stabile) o “fuori controllo” (OC, ossia presenta variazioni non casuali che richiedono indagini e azioni correttive).

In questo elaborato verrà presentata una carta di controllo denominata Distribution-Free chart based on the Score test, DFS in breve, che è stata appositamente studiata per monitorare contemporaneamente la media e la varianza di un processo. Questa tipologia di CC risulta particolarmente conveniente nelle frequenti situazioni in cui si è interessati a monitorare sia la media che la varianza di una certa distribuzione; in questo modo si può optare per un'unica carta invece di utilizzarne due, limitando così il problema della gestione dei falsi allarmi quando si utilizzano due CC in parallelo. Un’altra particolarità della carta DFS è il fatto che si tratta di una carta non parametrica; grazie a questa proprietà, applicare la carta alle situazioni più disparate sarà molto semplice, in quanto non vengono fatte assunzioni distributive sulla vera distribuzione del processo che andremo a controllare.

Presenteremo anche i risultati di un studio di simulazione volto a confrontare le performance della carta DFS con altre carte preesistenti. I risultati ottenuti sembrano notevoli, proprio perché DFS riesce a segnalare velocemente

situazioni di fuori controllo (OC), nonostante la sua natura non parametrica, qualunque sia la vera distribuzione in controllo (IC) dei dati.

Il seguente elaborato è diviso in quattro capitoli. Nel primo è presentata una introduzione agli strumenti del CSQ e alla terminologia utilizzata in questo mondo. Nel secondo viene presentata formalmente la carta DFS. Nel terzo vengono presentate le performance della carta. Nell'ultimo capitolo, la carta DFS viene implementata in R (linguaggio di programmazione) e applicata a due dataset reali, oltre a dei casi in cui vengono simulate delle situazioni OC per verificare quanto sia sensibile la carta.

Un possibile dubbio che potrebbe sorgere riguardo questa carta di controllo è il seguente: "*Ma quando questa carta chiama un allarme, lo sta facendo perché è andata fuori controllo la media o la varianza? Oppure entrambe?*". Io me lo sono posto questo dubbio, e dato che gli Autori (Ding D, Li J, Tsung F, Li Y (2023)) dell'articolo in cui viene presentata la carta DFS non hanno discusso l'argomento, ho deciso di cercare di dare una risposta a questa domanda. Il mio tentativo per risolvere questo problema lo troverete nel quarto capitolo di questa tesi.

CAPITOLO 1:

Le basi del Controllo Statistico della Qualità

1.1) Aspetti storici del Controllo Statistico della Qualità

Gli approcci statistici per controllare e migliorare la qualità di un processo, risalgono almeno agli anni 20 o 30 del 1900 con la pubblicazione del testo "Economic Control of Quality of Manufactured Product" (1931) di Walter A. Shewhart. Fin da allora il concetto di qualità e soprattutto il suo miglioramento, sono stati il focus di quel ramo della statistica che prende il nome di Controllo Statistico di Processo (in inglese SPC: Statistical Process Control).

Shewhart è conosciuto per essere stato il padre del Controllo Statistico della Qualità, è diventato famoso per aver elaborato la teoria delle cause comuni e di quelle speciali di varianza che agiscono sul processo. Gran parte della sua carriera professionale la trascorse alla Western Electric e nei laboratori della Bell Telephone. Alla Western Electric, prima dei lavori di Shewhart, venivano semplicemente controllati i prodotti finiti allo scopo di rimuovere quelli difettosi. Shewhart introdusse l'utilizzo di uno strumento grafico semplice ma molto efficace per controllare tutto il processo produttivo, allo scopo di ridurne la varianza e limitarne i difetti: le carte di controllo.

I controlli basati su criteri statistici ebbero la massima applicazione durante la seconda guerra mondiale, quando per l'industria bellica diventò necessario utilizzare in modo massiccio manodopera non specializzata e quindi soggetta ad un margine di errore maggiore. Durante il secondo dopoguerra, si iniziò a

parlare di qualità in modo sistematico grazie al Giappone, il quale dovette trovare un modo per riprendersi dalla profonda crisi economica nella quale si stava imbattendo dopo la sconfitta e che rappresentasse una nuova strada per essere competitivi nel mercato. La qualità per i giapponesi divenne uno strumento di rivalsa davanti al mondo, che permise di generare prodotti migliori a costi inferiori.

In tempi più recenti i metodi per il controllo della qualità sono diventati fondamentali per diverse attività industriali, aziendali e sanitarie, proprio perché un processo di qualità maggiore implica un minor numero di prodotti difettosi con un conseguente risparmio in termini di tempo, denaro e sforzo.

1.2) Il concetto di Qualità

La qualità di un prodotto o di un servizio può essere valutata secondo diversi aspetti, quali: prestazione, affidabilità, durata, manutenibilità, estetica, funzionalità, percezione del cliente e conformità alle normative vigenti, si capisce dunque come la qualità sia un'entità multiforme.

La tradizionale definizione di qualità ovvero: "*Qualità significa essere appropriato per l'uso*" purtroppo non ci è molto utile, in quanto non è facilmente quantificabile o applicabile dal punto di vista operativo; in ambito statistico si utilizza invece la seguente definizione: "*La qualità è inversamente proporzionale alla variabilità*". Basandoci su questo approccio, possiamo migliorare la qualità di un prodotto o di un servizio fornito, indagando sulle fonti di variabilità che influenzano il processo produttivo; esse si dividono convenzionalmente in comuni e speciali. Le cause comuni di variabilità sono quelle intrinseche del processo e sono ineliminabili. Le fonti di variabilità che non sono riconducibili a cause comuni, sono dunque eliminabili o almeno riducibili e prendono il nome di cause speciali di variabilità.

Se in un certo istante agiscono solamente le fonti comuni di variabilità, allora il

processo si dice in controllo (IC) altrimenti, quando agisce anche una sola fonte speciale di variabilità, esso si dice fuori controllo (OC). Ne deriva che, il lavoro dello statistico in ambito SPC sia proprio quello di monitorare il processo attraverso strumenti grafici (carte di controllo) e analisi dei dati (retrospettive o prospettiche), in modo da capire il prima possibile se nel processo stanno influendo anche fonti di variabilità speciale, al fine di proporre soluzioni per eliminarle e per tornare allo stato di processo IC, per garantire un adeguato livello di riproducibilità della qualità del prodotto/servizio.

1.3) Carte di Controllo

Le carte di controllo (CC) sono lo strumento grafico per eccellenza dell'SPC; esse tipicamente monitorano le variazioni di una statistica che rappresenta una caratteristica di qualità del processo al passare del tempo. Alla base delle CC (di fase I) sta un test statistico per verificare il sistema d'ipotesi in cui sotto l'ipotesi nulla, il processo è in controllo (IC) mentre sotto l'ipotesi alternativa il processo è fuori controllo (OC); quello che differenzia una CC da una banale verifica d'ipotesi, è il fatto che la prima ci suggerisce quando e come una fonte di variabilità speciale potrebbe aver agito sul processo.

$$H_0 : \left(\begin{array}{c} \text{il processo era in} \\ \text{controllo (stabile)} \\ \text{nel periodo in cui} \\ \text{sono stati raccolti} \\ \text{i dati} \end{array} \right) \text{ verso } H_1 : \left(\begin{array}{c} \text{il processo era} \\ \text{fuori controllo} \\ \text{(instabile) nel} \\ \text{periodo in cui} \\ \text{sono stati raccolti} \\ \text{i dati} \end{array} \right)$$

La forma standard di una carta di controllo è la seguente: si tratta di un grafico in cui nell'asse delle ascisse è riportato il numero del campione raccolto o l'istante di tempo (ordinato cronologicamente) e nell'asse delle ordinate è riportato il valore della statistica di controllo per ogni campione. Sono poi tracciate tre linee orizzontali: la Central Line (CL) che rappresenta il valore medio della statistica di controllo e i due limiti di controllo, uno superiore e uno

inferiore (rispettivamente UCL, Upper Control Limit e LCL, Lower Control Limit). Questi limiti vengono scelti in maniera tale che se il processo è IC, allora i valori della statistica di riferimento cadranno all'interno dei limiti con una probabilità molto elevata. Se in qualche istante la statistica di controllo fuoriesce dai limiti, viene segnalato un allarme e il processo è considerato OC; si dovrà quindi intervenire per individuare ed eliminare le cause speciali di variabilità. Un altro modo per cui è possibile accorgersi che il processo è andato fuori controllo, nonostante tutti i punti della statistica di controllo siano rimasti entro i limiti, è quello di verificare se sono presenti dei pattern sorprendenti nella CC. La soluzione più comune è quella di applicare delle regole supplementari come ad esempio le "Western Electric run rules", secondo le quali nella carta Shewhart $\bar{X}$, si segnala un allarme anche quando:

- Due punti tra tre consecutivi sono maggiori di $\hat{\mu}_0+2\hat{\sigma}_0/\sqrt{n}$ o minori di $\hat{\mu}_0 - 2\hat{\sigma}_0/\sqrt{n}$
- Quattro punti tra cinque consecutivi sono maggiori di $\hat{\mu}_0 + \hat{\sigma}_0/\sqrt{n}$ o minori di $\hat{\mu}_0 - \hat{\sigma}_0/\sqrt{n}$
- Otto punti consecutivi sono tutti sopra o tutti sotto $\hat{\mu}_0$

1.3.1) Carte di fase I e di fase II

Le carte di controllo si possono suddividere in due tipologie: quelle di fase I e quelle di fase II. Nelle CC di fase I si raccolgono e si analizzano retrospettivamente i dati del processo al fine di rispondere alla domanda: "*È cambiato qualcosa nel processo produttivo, durante il periodo osservato?*". Se il processo è rimasto IC (o altrimenti dopo aver eliminato con cognizione di causa i campioni responsabili della chiamata di un allarme), si stimano dei limiti di controllo adatti a monitorare i campioni futuri.

Nelle carte di fase II invece, ci chiediamo se il processo è cambiato, almeno fino all'ultimo istante di tempo osservato, e continuiamo a porci questa domanda ad ogni nuovo campione rilevato. Le CC di fase II sono dunque di tipo prospettico,

ovvero man mano che i dati vengono raccolti, disegniamo la statistica di controllo per ogni campione proveniente dal processo e la confrontiamo con i limiti di controllo stimati in fase I, al fine di individuare il più rapidamente possibile scostamenti dallo stato di processo IC.

1.3.2) FAP, ARL ed EDD

La scelta dei limiti di controllo per una CC non è assolutamente banale. Essi devono essere sufficientemente grandi da fare in modo che la probabilità di chiamare un falso allarme (false alarm probability, *FAP*) quando il processo è IC sia piccola, in genere 0.001. Al contempo i limiti devono essere abbastanza piccoli per accorgersi velocemente di una situazione OC. Per raggiungere questi obiettivi, una soluzione comune (proposta e utilizzata da Shewhart) è quella di porre i limiti a ±3 volte sigma (con sigma una stima dello scarto quadratico medio della statistica di controllo); questa soluzione, sotto assunzione di normalità della statistica, porta ad una probabilità di errore di I tipo pari a FAP=0.0027.

Rispondere alla domanda "*qual è una FAP accettabile per il mio processo?*" può rivelarsi poco immediato o intuitivo, per questo a volte risulta più comodo chiedersi: "*ogni quanto tempo sono disposto a interrompere il processo per un falso allarme?*"; per rispondere a questa domanda si ricorre alla Run Length (RL), definita come numero di istanti di tempo tra l'inizio della sorveglianza e il primo allarme. In particolare, la distribuzione della Run Length, è utile per valutare le performance di una carta di controllo. Un modo per sintetizzare una distribuzione è quella di considerare la sua media, definiamo dunque le seguenti due quantità:

termine	*acronimo*	*definizione*
average run length	ARL	$E\{RL\}$
expected detection delay	EDD	$E\{RL-\tau+1 \mid RL \geq \tau\}$

L'ARL è definito come il valore atteso della RL e vorremmo che esso sia grande

(+∞ ipoteticamente) quando il processo è IC. L'EDD invece è il numero di istanti medio che la carta impiega a chiamare un allarme a partire dall'istante in cui il processo si è sregolato (questo istante prende il nome di τ). A questo punto abbiamo definito gli strumenti base per confrontare l'efficienza di varie carte di controllo: a parità di ARL IC preferiremo una carta con EDD inferiore. Come vedremo alcune carte sono ottimali in certe situazioni e sono mediocri in altre; le carte di Shewhart ad esempio sono ottimali per situazioni di cambiamento isolato, oppure per grandi scostamenti dalla situazione IC, le carte CuSum invece sono ottimali per specifici "salti" (definiti in precedenza) del parametro del processo e sono meno efficienti per "salti" più grandi.

1.3.3) Carte di controllo Shewhart $\bar{X}$ e S per dati Normali

Le carte di controllo che tipicamente si studiano per prime in SPC sono le carte Shewhart $\bar{X}$ (di fase I), basate sulla media campionaria per controllare la posizione del processo e le Shewhart S (di fase I), basate sulla deviazione standard campionaria per controllare la scala del processo; inoltre queste 2 carte vengono spesso usate insieme per individuare eventuali variazioni sia in media che in varianza. Nella carta Shewhart $\bar{X}$ (di fase I) i limiti di controllo sono:

$$LCL = \hat{\mu}_0 - L\frac{\hat{\sigma}_0}{\sqrt{n}} \qquad \text{e} \qquad UCL = \hat{\mu}_0 + L\frac{\hat{\sigma}_0}{\sqrt{n}}$$

dove $\hat{\mu}_0$ è la media delle medie campionarie di tutti gli m campioni che abbiamo rilevato, mentre $\hat{\sigma}_0$ è la media delle deviazioni standard campionarie dei vari gruppi, divisa per un fattore di correzione $c_4(n-1)$ determinato in maniera tale che, sotto le nostre ipotesi $E\left(s_t/c_4(n-1)\right) = \sigma_0$. L invece è una costante che viene calcolata in modo tale da garantire una certa FAP desiderata. La scelta di L non è sempre semplice, in quanto allontanare i limiti di controllo dalla media, diminuirà l'errore di I tipo (segnalare un allarme quando in realtà il processo era in controllo) ma al contempo aumenterà l'errore di II tipo (dichiarare il processo come in controllo quando in realtà era fuori controllo) e viceversa, se invece si

decide di diminuire L. Una scelta che cerca di mettere d'accordo queste due tipologie di errore, deriva dalla diseguaglianza di Bonferroni, definendo L come il quantile di livello 1-(α/2m) di una Normale Standard. Un'altra soluzione può essere banalmente, come proposto da Shewhart, quella di usare L=3: fare ciò risulta sicuramente semplice, ma permette di ottenere una FAP piccola solo in casi particolari. La soluzione migliore è invece quella di calcolare L via simulazione, ciò garantirà la miglior efficienza possibile per la carta. Ecco le quantità appena descritte, mostrate in formule matematiche:

$$\hat{\mu}_0 = \text{Media}\left(\bar{x}_1, \ldots, \bar{x}_m\right) = \bar{\bar{x}} = \frac{1}{m}\sum_{t=1}^{m}\bar{x}_t = \frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n}x_{t,i}$$

con:

$$s_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{t,i} - \bar{x}_t\right)^2$$

$$\hat{\sigma}_0 = \text{Media}\left(s_1, \ldots, s_m \text{ "corretti"}\right) = \frac{1}{m}\sum_{t=1}^{m}\frac{s_t}{c_4(n-1)}$$

Nella carta Shewhart S (di fase I) i limiti di controllo superiore e inferiore sono:

$$LCL = \hat{\sigma}_0 - L\hat{\sigma}_0\sqrt{\frac{1 - c_4^2(n-1)}{c_4^2(n-1)}}$$

$$UCL = \hat{\sigma}_0 + L\hat{\sigma}_0\sqrt{\frac{1 - c_4^2(n-1)}{c_4^2(n-1)}}$$

le quantità che appaiono in queste due formule sono le stesse che abbiamo già incontrato, con la sola particolarità che L dovrà essere definita in maniera opportuna da ottenere la FAP accettabile. Un'altra formulazione dei limiti di controllo può essere: $LCL = L_1\hat{\sigma}_0$ e $UCL = L_2\hat{\sigma}_0$; utilizzando questa formulazione si tiene conto della asimmetria della distribuzione degli S_t ed è quindi preferibile.

Come già anticipato, l'uso di carte Shewhart $\bar{X}$ e S (di fase I) è quasi sempre combinato perchè spesso siamo interessati ad individuare sia cambiamenti in posizione che in dispersione; una seconda motivazione invece, è il fatto che per disegnare una carta di controllo per la sorveglianza prospettica (quindi di fase II), vogliamo identificare un sottoinsieme dei dati raccolti che sia rappresentativo del processo in controllo (IC). La procedura da seguire per ottenere un sottoinsieme rappresentativo dei dati in controllo, consiste quindi nel seguente algoritmo:

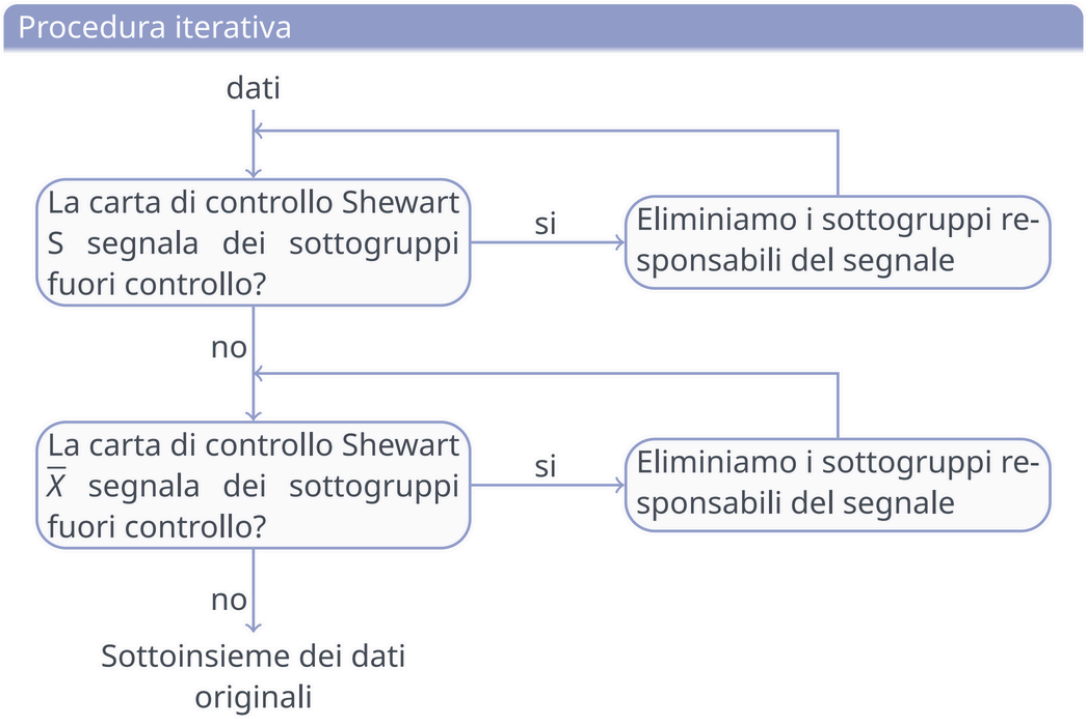


Figura 1.1 procedura iterativa per individuare un sottoinsieme idoneo a rappresentare il processo in controllo, per il disegno di una carta di fase II

È da notare il fatto che si parta applicando ai dati la carta Shewhart S (di fase I), in quanto essa non fa assunzioni particolari (oltre alla Normalità) e inoltre, dopo averla applicata, viene verificata l'omoschedasticità tra i gruppi (ovvero che le varianze dei gruppi siano simili). Proprio l'omoschedasticità è un'assunzione necessaria per applicare Shewhart $\bar{X}$ (di fase I). Possiamo poi proseguire iterativamente fino a che non otteniamo un sottoinsieme dei dati originari che non segnali allarmi, stando attenti a non eliminare "alla cieca" i sottogruppi responsabili dell'allarme: fare ciò porterebbe ad una sottostima della variabilità del processo e ad una conseguente sovrastima della capacità dello stesso.

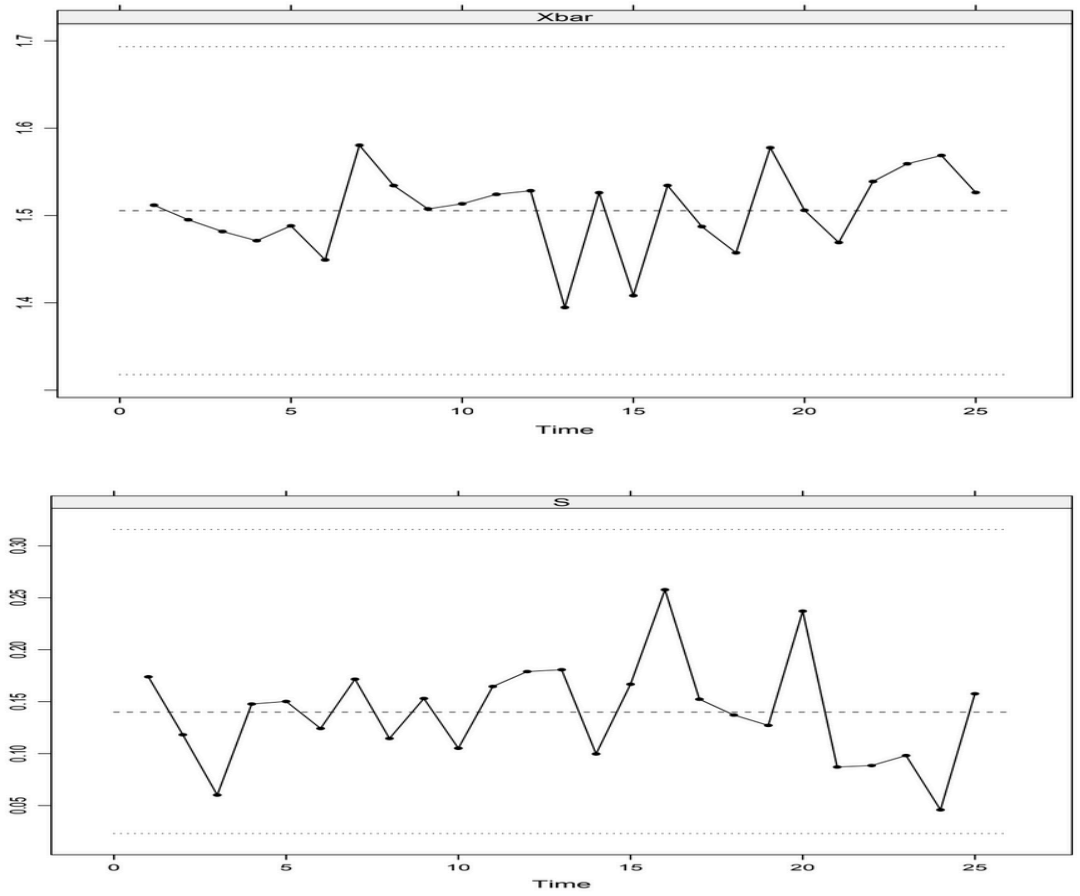


Figura 1.2 esempio di Shewhart $\bar{X}$ e Shewhart S (di fase I)

Le carte di controllo alla Shewhart sono semplici da usare e da interpretare, inoltre sono facilmente estendibili ad altri contesti, ad esempio, proporzioni (modello di riferimento: Binomiale), numero di difetti (modello di riferimento: Poisson), misure multivariate, profili/funzioni, dati complessi (ad esempio immagini). Rispetto ad altri approcci, le carte di Shewhart forniscono ottime prestazioni quando le cause speciali producono cambiamenti isolati nel tempo e di dimensione medio/grande; però sono meno efficienti di procedure alternative per identificare andamenti con una certa struttura, quali cicli, trend, ecc. Una soluzione può essere banalmente quella di usare delle procedure specifiche e più efficienti per scostamenti di piccola dimensione (ad esempio carte di tipo CuSum o EWMA), resta il fatto che è molto importante guardare l'andamento grafico della nostra carta e magari farsi aiutare da regole supplementari.

1.3.4) Carte di controllo EWMA

Le carte di controllo EWMA (Exponentially Weighted Moving Average) sfruttano il concetto di media mobile con pesi esponenziali o geometrici (o decrescenti in maniera esponenziale o geometrica) per definire una statistica di controllo del tipo:

$$W_t = \begin{cases} \mu_o & t = 0 \\ \lambda \bar{x}_t + (1-\lambda) W_{t-1} & t = 1, 2, \dots \end{cases} \quad \text{dove } 0 < \lambda \leq 1$$

Dove λ è la costante di lisciamento e μ_0 è il valore della media IC del processo; spesso μ_0 non è noto e viene stimato attraverso la media campionaria di un sottoinsieme di dati rappresentativi del processo IC, ottenuti con la procedura iterativa sopra descritta. La forza di questa tipologia di CC sta nel fatto che una media locale (pesata esponenzialmente), porta a dei dati con una minore variabilità rispetto ai dati originari, senza una grossa perdita di informazione.

Le carte di controllo EWMA risultano più efficaci delle carte Shewhart per individuare piccoli salti del livello del processo (o meglio per salti di ampiezza a cui si è deciso, durante la costruzione della carta, di dedicare particolare attenzione). Si è dimostrato che le carte EWMA (di fase II) e le CuSum (di fase II) hanno efficienza (valutata in termini di ARL OC) massima ed equiparabile tra loro, a parità di salto minimo che si vuole individuare. Questa efficienza si riesce a ottenere anche grazie al fatto che entrambe le tipologie di carte sono “con memoria”, ciò significa che ad ogni istante di tempo t, la statistica di controllo non è calcolata solamente sulle osservazioni del tempo t, bensì vengono utilizzate anche le statistiche di controllo degli istanti precedenti (t-1, t-2,...), correttamente pesate in modo da dare loro la giusta importanza. Viene fatto ciò, affinché la statistica di controllo sia una media locale ma rappresentativa dello stato del processo.

La varianza della statistica di controllo Wt è così definita:

$$L\frac{\sigma_0}{\sqrt{n}}\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2t}\right]}$$

Essa varia al variare della costante di lisciamento λ, in particolare: se λ aumenta, allora aumenta il peso dato alle ultime statistiche di controllo calcolate e di conseguenza la media diventerà più locale, ma allo stesso tempo aumenterà la variabilità del processo e di conseguenza i limiti di controllo, che sono così definiti:

$$\mu_0 \pm L\sqrt{\mathrm{var}(W_t)} = \mu_0 \pm L\frac{\sigma_0}{\sqrt{n}}\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2t}\right]}$$

Come si può notare, dopo un certo tempo t abbastanza grande, il termine $\left[1-(1-\lambda)^{2t}\right]$ si può approssimare ad 1, si ottengono così i nuovi LCL e UCL semplificati: $\mu_0 \pm L\lim_{t\to\infty}\sqrt{\mathrm{var}(W_t)} = \mu_0 \pm L\frac{\sigma_0}{\sqrt{n}}\sqrt{\frac{\lambda}{2-\lambda}}$

I parametri per la definizione delle carte di controllo EWMA sono quindi: il valore critico L e la costante di lisciamento λ, i quali vanno scelti in modo da avere prestazioni di ARL OC accettabili. Per calcolare queste 2 quantità si procede come segue: si fissano l'ARL in controllo, l'ARL fuori controllo e la dimensione del salto che si vuole individuare più rapidamente; si passa poi alla stima di L e λ. È utile ricordare che le carte EWMA non funzionano in modo eccellente per salti di grande dimensione, situazione in cui al contrario le Shewhart eccellono, per questo motivo queste due tipologie di carte sono spesso usate in modo combinato. Un altro aspetto da sottolineare è il fatto che anche le carte di tipo CuSum, basate sul log-rapporto di verosimiglianza, sono valide alternative alle EWMA. Dal punto di vista dell'ARL OC, le CuSum hanno prestazioni ottimali, questo perché si basano su un risultato teorico rigoroso, tratto da una versione del "Lemma fondamentale di Neyman-Pearson", esse però hanno un difetto: la difficile interpretazione di cosa stia "calcolando" la carta nei vari istanti di tempo, per questo motivo spesso si opta per le più intuitive EWMA.

1.3.5) Carte di controllo RS/P

Le carte di controllo RS/P (Recursive Segmentation and Permutation) sono un primo esempio di carte di controllo *non parametriche*, ecco alcune loro caratteristiche:

1) sono utilizzabili sia per dati individuali che in presenza di sottogruppi

2) sono utili per segnalare sia cambiamenti in posizione (media/livello) che in dispersione (variabilità/scala)

3) sono ragionevolmente efficienti sia verso cambiamenti isolati che verso cambiamenti più strutturati (come comportamenti ciclici, degradazioni progressive, interventi in tempi differenti di “cause speciali” con effetti permanenti,….)

4) Non richiedono informazioni a priori sulla distribuzione in controllo

Le carte di controllo non parametriche risultano molto utili quando non conosciamo bene la distribuzione in controllo del processo produttivo; siccome fare assunzioni distributive sbagliate porta a risultati per nulla attendibili, delle volte è meglio utilizzare carte non parametriche e lasciare che siano i dati a parlare.

Una delle caratteristiche fondamentali di RS/P è che per costruzione, la probabilità di segnalare un cambiamento di livello o dispersione quando il processo è IC, e quindi sbagliando, è sempre uguale al valore α prescelto, qualsiasi sia la dimensione del campione e qualsiasi sia la vera distribuzione del processo.

Vediamo ora un esempio di carta RS/P:

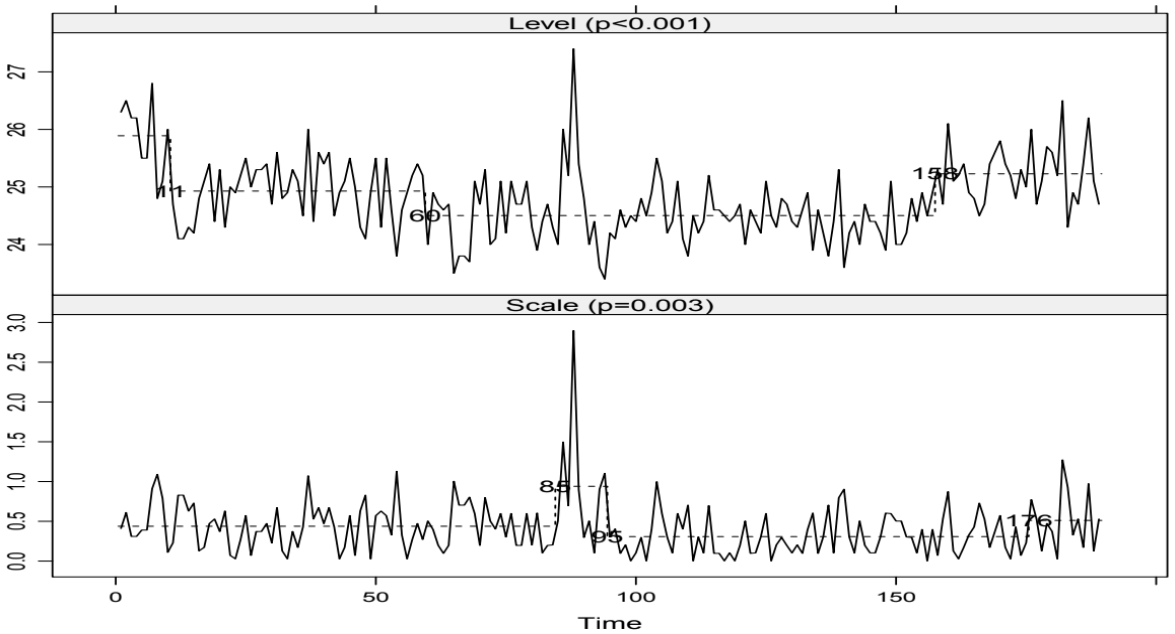


figura 1.3 Esempio di Shewhart RS/P

Da questa immagine possiamo capire come sia strutturata una carta RS/P, essa è composta da due pannelli; quello superiore riguarda possibili variazioni nel livello, e viene fornito il p-value per verificare il seguente sistema di ipotesi:

$$H_0 : \begin{pmatrix} \text{il livello medio è} \\ \text{stato costante nel} \\ \text{periodo di} \\ \text{osservazione} \end{pmatrix} \text{ verso } H_1 : \begin{pmatrix} \text{ci sono state delle} \\ \text{variazioni nel} \\ \text{livello medio} \end{pmatrix}$$

La linea continua mostra le medie dei sottogruppi (o, nel caso di dati individuali, le osservazioni stesse). La linea tratteggiata mostra una stima (eventualmente variabile nel tempo) del livello medio, può essere o una costante o una funzione a gradini e viene usata a fini diagnostici per capire dove e come è cambiato il livello medio.

Il pannello inferiore invece riguarda possibili variazioni nella scala, il sistema di ipotesi per il quale viene fornito il p-value riportato nel “titolo” del pannello è il seguente:

$$H_0 : \begin{pmatrix} \text{la dispersione è stato} \\ \text{costante nel periodo} \\ \text{di osservazione} \end{pmatrix} \text{ verso } H_1 : \begin{pmatrix} \text{ci sono state delle} \\ \text{variazioni nella} \\ \text{dispersione} \end{pmatrix}$$

La linea continua mostra una misura della dispersione, così definita:

$$\sqrt{n^{-1} \sum_{i=1}^{n} \left(x_{t,i} - \hat{\mu}_t\right)^2}$$

dove $\hat{\mu}_t$ è la stima del livello, ovvero la linea tratteggiata del pannello superiore, mentre in questo pannello, la linea tratteggiata mostra una stima (eventualmente variabile nel tempo) della dispersione media.

Il concetto che sta alla base di RS/P consiste nel calcolare K+1 statistiche test elementari, partendo da T_0 che ipotizza cambiamenti isolati (come per la Shewhart $\bar{X}$), si prosegue poi con T_1, che ipotizza *un* "salto nella media", l'istante del "salto" è determinato dai dati, si continua poi con T_2 che ipotizza *due* "salti nella media" e così via. Una volta calcolate queste statistiche si cerca quella più sorprendente per un processo IC, in che modo? Definendo:

$$G_{oss} = \max_{i=0,\ldots,K} \frac{T_i - \mathrm{E}_{IC}\left(T_i\right)}{\sqrt{\mathrm{var}_{IC}\left(T_i\right)}}$$

Si tratta di una statistica test che combina tutte le precedenti, in modo da poter verificare H_0 contro tutte le possibili H_1 delle K+1 statistiche elementari; resta però ancora un problema: non conosciamo la distribuzione di G, la quale ci serve per calcolare il p-value: $p = \Pr_{IC}\left(G > G_{oss}\right)$

Per risolvere questo problema è necessario utilizzare un approccio di permutazione, sfruttando questo risultato chiave:

Supponiamo che il processo sia in controllo e che quindi tutte le osservazioni siano i.i.d. con funzione di densità $p(\cdot)$, ovvero, $x_{t,i} \sim p(\cdot)$ per qualsivoglia t e i. Indichiamo con $Y = (y_1, \ldots, y_N) = (x_{1,1}, \ldots, x_{m,n})$ il vettore di tutte le $N = n \cdot m$

osservazioni e con $Y_{()} = (y_{(1)}, \ldots, y_{(N)})$ il corrispondente vettore delle statistiche ordinate $y_{(1)} < y_{(2)} < \cdots < y_{(N)}$ Allora:

$$\Pr\left(\boldsymbol{Y} = \boldsymbol{a} | \boldsymbol{Y}_{()}\right) = \begin{cases} \frac{1}{N!} & \text{se } \boldsymbol{a} \text{ è una permutazione di } \boldsymbol{Y}_{()} \\ 0 & \text{altrimenti} \end{cases}$$

In sostanza, riusciamo ad avere la distribuzione condizionata, anche se non conosciamo la distribuzione di partenza, e ciò ci permette comunque di calcolare il valore atteso e la varianza IC delle statistiche test, solo che computazionalmente risulterebbe molto oneroso farlo per tutte e N! permutazioni; si passa dunque ad un approccio per simulazione di tipo bootstrap. Riusciamo così ad ottenere un errore di primo tipo α del valore esatto prestabilito. In conclusione è importante ricordare che RS/P è una carta di controllo non parametrica con discrete e soddisfacenti performance, ma mai ottimali, qualunque sia la distribuzione effettiva del processo o la numerosità campionaria dei singoli gruppi.

CAPITOLO 2: La carta di controllo DFS: Distribution-Free chart based on the Score test

2.1) Introduzione alle carte non parametriche per il monitoraggio di posizione e scala

In questo capitolo vi presenterò la nuova carta di controllo DFS, sviluppata da Fugee Tsung e Yang Li (2023) si tratta, come si può capire dalla denominazione, di una carta *non parametrica* basata sullo score di verosimiglianza per sorvegliare *congiuntamente* posizione e scala. La particolarità di questa carta sta proprio nel fatto che si prefigge come obiettivo quello di individuare congiuntamente queste situazioni di fuori controllo, al contrario di altre carte (sempre non parametriche) che monitorano solo la media o solo la varianza singolarmente. Esempi di CC non parametriche per controllare la media del processo risalgono a Bakir e Reynolds (1979), con una carta che si basa sulla statistica dei ranghi con segno di Wilcoxon, Hackl e Ledolter (1991) con una EWMA basata sui ranghi standardizzati. Più recentemente, Chakraborti et al. (2009) hanno sfruttato il test di precedenza per carte di Fase II, Graham et al. (2011-2012) hanno proposto carte di controllo basate sui ranghi con segno e su test sequenziali, Liu et al. (2014) hanno progettato uno schema ad avviamento automatico basato su ranghi sequenziali; di questa tipologia di carte ne esistono molte altre, ma queste erano tra le più significative.

Tutti questi esempi ci fanno capire quanto siano importanti e sempre più necessarie, carte di controllo non parametriche; il limite delle carte sopra citate sta nel fatto che esse si occupano solamente di individuare scostamenti in media e non si occupano di monitorare la varianza. A questo scopo sono stati introdotti metodi non parametrici basati sui ranghi per monitorare congiuntamente livello e scala. Vediamo degli esempi di queste carte: Mukherjee e Chakraborty (2012) hanno costruito una carta basata sulla combinazione delle statistiche di Wilcoxon e Ansari-Bradley, per il monitoraggio di posizione e scala, rispettivamente. Ross e al. (2011) hanno combinato le statistiche di Mann-Whitney e di Mood. Chowdhury (2014) e Xiang (2019) hanno creato una carta di controllo basata sulla statistica di Cucconi, la quale è in realtà la combinazione di due statistiche di rango.

Un altro approccio al problema è quello che considera l'utilizzo di *test di bontà di adattamento* (Goodness Of Fit, GOF) per rilevare cambiamenti arbitrari. Importanti riconoscimenti vanno fatti a Zou e Tsung (2010) per il loro lavoro al fine di integrare metodi GOF nel mondo del SPC. Altri riconoscimenti vanno a Ross & Adams (2012) e Zhang (2017) per il loro lavoro nei metodi GOF. Si può vedere come quasi tutte le procedure basate sui ranghi combinino due statistiche separate in qualche modo artificiale, ad esempio facendone la somma dei quadrati o il valore massimo. Tuttavia combinare in modo così semplice e arbitrario delle statistiche non garantisce assolutamente l'ottimalità della carta di controllo. Inoltre, molti metodi GOF richiedono che la dimensione del campione sia maggiore di uno, ciò riduce la rosa di casi in cui questi metodi possono essere utilizzati.

Per superare queste carenze, viene proposto un nuovo metodo non parametrico per monitorare congiuntamente posizione e scala, basato su un risultato rigoroso, ovvero la carta di controllo DFS: Distribution-Free chart based on the Score test. DFS si basa sullo score di verosimiglianza per l'inferenza dei parametri di posizione e scala. Con un'opportuna trasformazione, la statistica

test parametrica basata sullo Score di verosimiglianza, si trasforma in una statistica non parametrica, che non richiede la conoscenza della distribuzione del processo; inoltre essa è adatta sia per osservazioni individuali che per gruppi. Le simulazioni mostrano che la carta DFS è molto efficiente per rilevare cambiamenti nella posizione e/o nella scala, inoltre essa si è rivelata una carta robusta sotto varie distribuzioni del processo.

2.2) Introduzione alla nuova carta DFS: Distribution-Free chart based on the Score test, la formulazione della carta e le difficoltà incontrate

La carta DFS che verrà proposta è di fase II, come sappiamo, ciò significa che i dati ci arrivano man mano che il tempo scorre e ad ogni istante di tempo i nuovi dati vengono analizzati, viene calcolata la statistica di controllo e confrontata con i limiti, se viene segnalato un allarme, allora il processo viene interrotto per capire cosa non ha funzionato e soprattutto come riportarlo IC. Iniziamo ora con la presentazione del problema e della carta: assumiamo che i dati del processo seguano una distribuzione continua con funzione di ripartizione $G(x; \mu, \sigma) = F((x - \mu)/\sigma)$, con μ parametro di posizione, $\sigma > 0$ parametro di scala, e $F(\cdot)$ funzione di ripartizione della distribuzione standardizzata, cioè con $\mu = 0$ e $\sigma = 1$. La corrispondente funzione di densità risulta quindi $g(x; \mu, \sigma) = (1/\sigma) * f((x - \mu)/\sigma)$.

Durante la fase II, il nostro compito è quello di verificare il più velocemente possibile se il parametro di posizione o di scala cambiano, dunque alla base della carta sta il seguente sistema di verifica d'ipotesi:

$$H0 : \mu = \mu_0 \text{ e } \sigma = \sigma_0 \qquad \text{rispetto a} \qquad H1 : \mu \neq \mu_0 \text{ o } \sigma \neq \sigma_0$$

dove μ_0 e σ_0 sono i parametri quando il processo è IC.

Per calcolare lo Score test abbiamo bisogno delle seguenti due funzioni score:

$$\frac{\partial \ln g(x;\mu,\sigma)}{\partial \mu} = -\frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \Big/ f\left(\frac{x-\mu}{\sigma}\right),$$

$$\frac{\partial \ln g(x;\mu,\sigma)}{\partial \sigma} = -\frac{1}{\sigma}\left[1 + \frac{x-\mu}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \Big/ f\left(\frac{x-\mu}{\sigma}\right)\right].$$

L'informazione di Fisher è quindi espressa come:

$$\mathbf{I}(g) = \frac{1}{\sigma^2}\mathbf{I}(f) = \frac{1}{\sigma^2}\begin{bmatrix} I_{11}(f) & I_{12}(f) \\ I_{12}(f) & I_{22}(f) \end{bmatrix}$$

Dove:

$$I_{11} = \int_{-\infty}^{\infty} \left(\frac{f'(y)}{f(y)}\right)^2 f(y)\mathrm{d}y,$$

$$I_{12} = \int_{-\infty}^{\infty} \left(\frac{f'(y)}{f(y)}\right)\left(1 + \frac{f'(y)}{f(y)}\right) f(y)\mathrm{d}y,$$

$$I_{22} = \int_{-\infty}^{\infty} \left(1 + \frac{f'(y)}{f(y)}\right)^2 f(y)\mathrm{d}y.$$

Pertanto, dato un campione di dimensione N (con N ≥ 1) costituito da osservazioni $x_1,...,x_n$, una statistica per lo Score test prende tipicamente la seguente forma:

$$\frac{1}{N}\mathbf{s}^T \mathbf{I}^{-1}(f)\mathbf{s}, \quad (1)$$

dove $\mathbf{s} = [s_1, s_2]^T$ con:

$$s_1 = -\sum_{i=1}^{N} f'\left(\frac{x_i-\mu_0}{\sigma_0}\right) \Big/ f\left(\frac{x_i-\mu_0}{\sigma_0}\right),$$

$$s_2 = -\sum_{i=1}^{N} \left(1 + \frac{x_i-\mu_0}{\sigma_0} f'\left(\frac{x_i-\mu_0}{\sigma_0}\right) \Big/ f\left(\frac{x_i-\mu_0}{\sigma_0}\right)\right).$$

Chiaramente, per calcolare la statistica, oltre ai parametri μ_0 e σ_0, i quali solitamente vengono stimati nella Fase I, sono necessarie le forme specifiche di

$f(\cdot)$ e $f'(\cdot)$, ed è proprio la loro stima la principale difficoltà nella costruzione della statistica di controllo.

2.2.1) La statistica di controllo della carta DFS

Individuare la statistica dello score test dell'espressione (1) richiede una procedura molto onerosa dal punto di vista computazionale, per non parlare del fatto che $f(\cdot)$ e $f'(\cdot)$ sono ignote. Per affrontare il problema, faremo prima qualche trasformazione in modo da semplificare la statistica. Ponendo $z_i = (x_i - \mu_0)/\sigma_0$ come valore standardizzato, allora esso può anche essere riscritto come l'inverso della sua funzione di ripartizione, ovvero $z_i = F^{-1}(u_i)$ per $0 < u_i < 1$. Oppure in modo equivalente, z_i può essere visto come un quantile e ui come il valore della distribuzione: $u_i = G(x_i; \mu_0, \sigma_0) = F((x_i - \mu_0)/\sigma_0)$ per l'osservazione x_i. Ora andiamo a definire le seguenti due funzioni:

$$\phi_1(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))},$$

$$\phi_2(u) = -1 - F^{-1}(u)\frac{f'(F^{-1}(u))}{f(F^{-1}(u))},$$

dove $F^{-1}(u)$ è un quantile, ovvero $F^{-1}(u) = \inf\{y : F(y) \geq u\}$, allora le funzioni score (s_1 e s_2) possono essere riscritte come:

$$s_1 = \sum_{i=1}^{N} \phi_1(u_i), \quad s_2 = \sum_{i=1}^{N} \phi_2(u_i),$$

e di conseguenza la matrice d'informazione di Fisher avrà le seguenti componenti:

$$I_{11}(f) = \int_0^1 \phi_1^2(u)\mathrm{d}u,$$

$$I_{12}(f) = \int_0^1 \phi_1(u)\phi_2(u)\mathrm{d}u,$$

$$I_{22}(f) = \int_0^1 \phi_2^2(u)\mathrm{d}u.$$

Grazie alle trasformazioni $\phi_1(\cdot)$ e $\phi_2(\cdot)$, la statistica nell'espressione (1) si può scrivere in modo più semplice, ma rimane il problema della forma di $f(\cdot)$ e di $f'(\cdot)$. Ne consegue che $f(\cdot)$ deve essere selezionata correttamente, in modo che la statistica (1) abbia una forma esplicita e soddisfacente capacità di segnalare allarmi. Per ricavare la statistica di controllo, gli Autori hanno proposto che $f(\cdot)$ sia la funzione di densità della distribuzione logistica standard, ovvero:

$$f(y) = \frac{\mathrm{e}^{-y}}{(1+\mathrm{e}^{-y})^2}.$$

così facendo, $\phi_1(u)$ e $\phi_2(u)$ diventano rispettivamente: $\phi_1(u) = 2u - 1$ e $\phi_2(u) = (2u - 1) * \ln(u/(1 - u)) - 1$

L'informazione di Fisher diventa invece

$$\mathbf{I} = \begin{bmatrix} 1/3 & 0 \\ 0 & (\pi^2+3)/9 \end{bmatrix}$$

Ci sono sostanzialmente due ragioni per cui viene scelta la distribuzione logistica, primo perché con questa scelta, la statistica di controllo ha prestazioni robuste qualunque sia la vera distribuzione del processo; ciò lo verificheremo nel capitolo 3, dove vedremo i risultati di uno studio di simulazione, nel quale sono stati passati alla carta DFS dati generati da diverse distribuzioni tipiche. In secondo luogo, la scelta della distribuzione logistica permette di assegnare a $\phi_1(\cdot)$ e $\phi_2(\cdot)$ un significato ben preciso: $\phi_1(\cdot)$ ha il compito di individuare eventuali scostamenti di livello, mentre $\phi_2(\cdot)$ è incaricato di individuare

eventuali scostamenti di scala. Nello specifico, la funzione $\phi_1(u) = 2u - 1$, se $u = F(\cdot)$ con $F(\cdot)$ funzione di ripartizione della distribuzione logistica standard; allora otteniamo esattamente la stessa forma del rango standardizzato. Il rango standardizzato proposto da Hackl e Ledolter è infatti $2F(x_i) - 1$ per un'osservazione x_i, ed è così poiché la funzione di ripartizione caratterizza semplicemente la posizione relativa o l'informazione sull'ordine di una certa osservazione x_i. In concreto, il metodo dei ranghi standardizzati di Hackl e Ledolter, è solamente una versione standardizzata del test della somma dei ranghi di Wilcoxon.

Secondo Hajek, nell'ambito delle variabili continue, la statistica test della somma dei ranghi di Wilcoxon, è il test basato sui ranghi asintoticamente più potente, per testare uno scostamento della media, se le osservazioni sono soggette a una distribuzione logistica. Ciò implica che la scelta della distribuzione logistica fatta in precedenza per ricavare le forme specifiche di $\phi_1(\cdot)$ e $\phi_2(\cdot)$ sia quantomeno una scelta naturale.

Al fine di applicare la carta DFS, assumiamo che la funzione di ripartizione quando il processo è IC, che indicheremo con $G_0(x)$, sia nota o sia stato possibile stimarla dalla funzione di ripartizione empirica durante la fase I. Di conseguenza, per il k-esimo campione costituito da osservazioni $x_{k1},\ldots,x_{kn}$, la statistica di controllo può ora essere scritta come:

$$Q_k = \frac{1}{N}\boldsymbol{\phi}_k^T \mathbf{I}^{-1}\boldsymbol{\phi}_k$$

dove:

$$\boldsymbol{\phi}_k = \left[\sum_{i=1}^{N}\phi_1(G_0(x_{ki})),\quad \sum_{i=1}^{N}\phi_2(G_0(x_{ki}))\right]^T$$

La statistica Q_k sopra definita ha forma di una tipica carta di Shewhart senza

memoria, che sfrutta solo le osservazioni al tempo t per calcolare la statistica di controllo e non quelle precedenti; inoltre essa non è sensibile a individuare piccoli scostamenti. Pertanto, al fine di rendere la carta DFS più sensibile a piccoli scostamenti, e farla diventare una carta "con memoria", è stato incorporato lo schema EWMA. Si sostituisce dunque $\boldsymbol{\phi}_k$ con la seguente quantità: $\boldsymbol{\theta}_k = (1-\lambda)\boldsymbol{\theta}_{k-1} + \lambda\boldsymbol{\phi}_k$

dove λ è il parametro di lisciamento, quindi ovviamente vale $0 < \lambda \leq 1$, infine la statistica di controllo che viene effettivamente tracciata nella carta di controllo è dunque:

$$R_k = \frac{1}{N}\boldsymbol{\theta}_k^T \mathbf{I}^{-1}\boldsymbol{\theta}_k \quad (2)$$

Vale la pena notare che per applicare la carta DFS, bisogna solo stimare la funzione di ripartizione quando il processo è IC, ovvero $G_0(\cdot)$, cosa che viene comunemente fatta nella Fase I grazie alla funzione di ripartizione empirica. Ne deriva che il calcolo della statistica di controllo R_k, è abbastanza semplice. Da ricordare inoltre, è il fatto che non stiamo facendo alcuna assunzione distributiva sul processo (ovvero sulla distribuzione di X). Infine è da sottolineare il fatto che DFS funziona anche nel caso di osservazioni individuali, la statistica grafica (2) infatti è ancora applicabile.

Arriviamo ora ai limiti della carta proposta, come più volte ribadito, essa può rilevare sia scostamenti di posizione che di scala, ma non riesce a essere più specifica: se viene chiamato un allarme, non si può sapere se esso è stato chiamato per una sregolazione della media, oppure per una sregolazione della varianza, oppure ancora per una combinazione delle due. Un'ultima cosa da ricordare è che alla base di DFS vi è una quantità statistica fondamentale: lo score di verosimiglianza, dunque le prestazioni di questa carta sono garantite da risultati teorici rigorosi.

CAPITOLO 3: Valutazione delle performance di DFS via simulazione

L'obiettivo di questo capitolo è quello di mostrare le performance teoriche della carta DFS. Riporterò ora i risultati presentati dagli Autori della carta DFS (Ding D, Li J, Tsung F, Li Y (2023)), ottenuti via simulazione, basati su $10^5$000 replicazioni. Le capacità della carta DFS verranno confrontate con altre tre carte, basate su metodi differenti. Per confrontare le prestazioni delle varie carte, è stato uniformato il valore di ARL IC, ponendolo pari a 370. Diremo quindi che una carta di controllo è migliore di un'altra se il suo ARL OC è minore; infatti se l'ARL OC è minore, significa che quella carta impiega meno istanti di tempo (o meno campioni rilevati) a segnalare un allarme, a partire dall'istante τ (istante in cui il processo passa da IC a OC).

3.1) La carta di controllo CEW

La carta di controllo CEW (Combinazione di due EWMA), è la combinazione di una classica EWMA per la media e di una EWMA unilaterale per la varianza. Date le osservazioni x_t, per t=1,2,... , nel caso in cui la media in controllo sia nulla, la carta CEW è basata sulle seguenti due statistiche:

$$w_t = (1-\lambda)w_{t-1} + \lambda x_t \quad \text{e} \quad v_t = (1-\lambda)\max\{1, v_{t-1}\} + \lambda x_t^2$$

La carta CEW è quindi parametrica. La scelta di una carta di tipo EWMA non è casuale: rispetto a una carta di tipo Shewhart, carte di tipo EWMA o CuSum, se ben applicate, risultano più robuste; viene scelto il primo tipo di carta per la sua semplice interpretazione e per uniformità con la carta DFS.

3.2) La carta di controllo NLE

La carta di controllo NLE (Non-parametric Likelihood-ratio EWMA) è di tipo non parametrico, si basa sul test di bontà di adattamento ed è stata proposta da Zou e Tsung. Questa carta risulta molto robusta ed efficiente per identificare cambiamenti in livello, scala e forma, per questo è molto interessante il suo confronto con DFS. La statistica di controllo di NLE é così definita:

$$z_t = (1-\lambda)z_{t-1} + \lambda y_t$$

dove:

$$y_t = \frac{1}{1-G_t^{(\lambda)}(x_t)} \ln\left(\frac{G_t^{(\lambda)}(x_t)}{G_0(x_t)}\right) + \frac{1}{G_t^{(\lambda)}(x_t)} \ln\left(\frac{1-G_t^{(\lambda)}(x_t)}{1-G_0(x_t)}\right)$$

con $G_0(\cdot)$ funzione di ripartizione di x_t quando il processo è IC; e con $G_i^{(\lambda)}(\cdot)$ funzione di ripartizione pesata così definita:

$$G_i^{(\lambda)}(u) = a_{\lambda,i}^{-1} \sum_{j=1}^{i} (1-\lambda)^{i-j} I_{\{X_j \le u\}}, \quad a_{\lambda,i} = \sum_{j=1}^{i} (1-\lambda)^{i-j}$$

per ogni punto u.

3.3) La carta di controllo WAB

La carta di controllo WAB (Wilcoxon Ansari-Bradley) è di tipo non parametrico, si basa sui ranghi ed è stata proposta da Mukherjee e Chakraborti. Essa è composta da due parti: la prima si basa sulla somma dei quadrati della statistica dei ranghi sommati di Wilcoxon per controllare la media; la seconda si basa sul quadrato della statistica test di Ansari-Bradley per controllare la varianza. In questa applicazione della carta carta WAB si è deciso di apportare alcune modifiche, ovvero lasciare che la dimensione del primo campione sia infinito e

che quella del secondo sia uno. In questo modo otteniamo delle statistiche test, per posizione e scala, che risultano standardizzate. Esse si presentano rispettivamente come:

$$\sqrt{3}(2G_0(x_t) - 1) \quad \text{e} \quad \sqrt{3}(4|G_0(x_t) - 0.5| - 1) \quad .$$

Al fine di rendere confrontabile questa carta con le altre, la statistica di controllo di tipo Shewhart viene adattata allo schema EWMA, in questo modo la statistica di controllo risulta:

$$3(2b_{t1} - 1)^2 + 3(4b_{t2} - 1)^2$$

dove: $b_{t1} = (1 - \lambda)b_{t-1,1} + \lambda G_0(x_t)$ con $b_{01} = 0.5$

e $b_{t2} = (1 - \lambda)b_{t-1,2} + \lambda|G_0(x_t) - 0.5|$ con $b_{02} = 0.25$

3.4) Confronto dei risultati ottenuti e commento

Al fine di valutare l'efficienza e la robustezza di ciascuna carta, si applicano le 4 carte di controllo a dati simulati da 3 tipiche distribuzioni. Esse sono: (i) la distribuzione Normale Standard $N(0,1)$; (ii) la distribuzione t di Student con 3 gradi di libertà $t(3)$; (iii) e la distribuzione chi quadro con 3 gradi di libertà $\chi^2(3)$. Le distribuzioni chi quadro e t di student sono inoltre standardizzate in modo da avere media nulla e varianza unitaria.

La seguente Tabella mostra gli ARL OC delle 4 carte con le osservazioni generate da una Normale, prima simulando un salto di δ nella media e poi simulando una sregolazione della varianza (che passa da 1 a δ^2). Per agevolare la lettura, la miglior performance (in termini di ARL), al variare di δ, è evidenziata in **grassetto**.

TABLE 1 OC ARL comparisons under normal distribution.

$N(0,1)$ versus $N(\delta,1)$					$N(0,1)$ versus $N(0,\delta^2)$				
δ	DFS	NLE	WAB	CEW	δ	DFS	NLE	WAB	CEW
0.00	369	371	370	370	1.00	370	369	371	370
0.25	123	120	110	**107**	1.10	**115**	127	179	131
0.50	37.2	37.7	35.2	**33.0**	1.20	**51.7**	58.4	95.8	60.8
0.75	17.5	19.1	17.9	**16.0**	1.30	**31.1**	33.4	60.1	34.3
1.00	10.7	12.2	11.8	**10.3**	1.40	**20.9**	22.4	42.9	22.6
1.50	5.67	6.54	7.45	**5.66**	1.60	**12.4**	12.8	26.8	12.5
2.00	3.73	4.01	5.89	**3.69**	1.80	8.72	8.82	19.7	**8.51**
3.00	2.14	**1.95**	4.90	1.99	2.00	6.75	6.66	16.1	**6.56**
4.00	1.49	**1.26**	4.72	1.32	3.00	3.43	3.19	10.2	**3.15**

Per quanto riguarda il controllo della media, la carta CEW è sicuramente la migliore, almeno per salti piccoli, medi e moderati, questo perchè CEW è una carta parametrica costruita proprio sulla distribuzione Normale. Al contempo, la carta DFS ha delle performance soddisfacenti ed è migliore sia di NLE che di WAB, almeno per salti da piccoli a moderati ($0.5 \leq \delta \leq 2$). In pratica per $\delta \geq 0.75$ le differenze tra DFS e CEW sono minime.

Per quanto riguarda il controllo della varianza, la carta DFS si rivela sorprendentemente essere la migliore, almeno per sregolazioni da piccole a moderate (per $\delta \leq 1.6$). Per scostamenti maggiori $\delta \geq 1.8$, CEW risulta la migliore, ma la differenza con DFS è trascurabile. Questo risultato sorprende in quanto raramente una carta non parametrica riesce a performare meglio di una parametrica (con corretta assunzione distributiva).

Un'altra cosa degna di nota è la scarsa efficacia di WAB in questa applicazione, soprattutto per salti della media di grande dimensione ($\delta \geq 1.5$); questo è dovuto alla natura di WAB: essa si basa sui ranghi e come è noto, essendo una statistica d'ordine, anche se dei valori di x raggiungono valori estremi, può succedere che il valore della statistica non cambi. WAB si rivela anche molto scarsa nell'individuare variazioni nella scala del processo, come vedremo sarà così anche per le altre due distribuzioni: chi quadro e t di student.

Prima di continuare con il commento dei risultati, è da far notare che i risultati della carta CEW non saranno molto utili, questo perchè applicare una carta

parametrica quando l'assunzione distributiva è sbagliata, porta a notevoli bias nell'ARL IC della stessa. In questa situazione CEW chiama troppi falsi allarmi. Come vedremo nelle seguenti tabelle, quando la situazione è IC, per la carta CEW l'ARL stimato sarà notevolmente inferiore a quello nominale, tanto che se volessimo riportarlo a 370, dovremmo utilizzare $\lambda \leq 0.001$; utilizzare un parametro di lisciamento così piccolo non è fattibile nella realtà. Per i motivi appena spiegati la colonna di CEW non è da considerare al fine di individuare la carta con performance migliore.

TABLE 2 OC ARL comparisons under t distribution.

$\frac{t_3}{\sqrt{3}}$ versus $\frac{t_3}{\sqrt{3}} + \delta$					$\frac{t_3}{\sqrt{3}}$ versus $\delta \cdot \frac{t_3}{\sqrt{3}}$				
δ	DFS	NLE	WAB	CEW	δ	DFS	NLE	WAB	CEW
0.00	370	371	370	108	1.00	370	369	370	108
0.25	89.8	86.7	**67.5**	81.9	1.10	**177**	220	207	78.7
0.50	23.7	26.8	**19.1**	28.6	1.20	**100**	132	123	59.0
0.75	11.9	15.4	**10.3**	16.6	1.30	**63.6**	85.5	79.8	44.4
1.00	8.00	10.9	**7.30**	10.4	1.40	**44.5**	60.3	56.4	35.6
1.50	**4.99**	6.86	5.09	5.93	1.60	**26.1**	34.4	34.8	23.6
2.00	**3.80**	4.89	4.35	3.93	1.80	**18.2**	23.2	25.0	16.5
3.00	**2.93**	3.14	4.06	2.06	2.00	**14.0**	17.6	19.6	12.6
4.00	**2.25**	2.35	4.02	1.29	3.00	**7.01**	8.10	11.1	5.49

Sotto distribuzione t di student, WAB risulta essere la carta migliore per salti della media di piccola dimensione $0.25 \leq \delta \leq 1$; quando i salti diventano di grande dimensione $1.5 \leq \delta \leq 4$ invece, la carta che eccelle è proprio DFS. Per quanto riguarda variazioni di scala, DFS è la migliore per qualunque salto considerato.

TABLE 3 OC ARL comparisons under χ^2 distribution.

$(\chi^2_3\text{-}3)/\sqrt{6}$ versus $(\chi^2_3\text{-}3)/\sqrt{6}+\delta$					$(\chi^2_3\text{-}3)/\sqrt{6}$ versus $\delta \cdot (\chi^2_3\text{-}3)/\sqrt{6}$				
δ	DFS	NLE	WAB	CEW	δ	DFS	NLE	WAB	CEW
0.00	371	370	370	101	1.00	371	372	370	100
0.25	211	**88.7**	107	61.7	1.10	23.8	**23.4**	122	66.6
0.50	43.4	**26.8**	29.1	30.9	1.20	11.2	**10.9**	57.3	46.9
0.75	18.5	18.4	**15.5**	17.0	1.30	7.47	**7.36**	35.9	34.7
1.00	11.9	14.4	**10.3**	11.0	1.40	5.71	**5.67**	26.4	26.1
1.50	7.15	9.82	**6.23**	6.23	1.60	4.10	**4.06**	17.8	16.9
2.00	5.16	7.12	**4.91**	4.17	1.80	3.36	**3.30**	14.0	11.7
3.00	**3.41**	4.22	4.00	2.16	2.00	2.94	**2.92**	11.9	9.01
4.00	**2.67**	2.76	4.00	1.39	3.00	2.10	**2.03**	8.24	3.80

Sotto distribuzione χ^2, ognuna delle tre carte risulta avere una dimensione del salto in cui eccelle ad individuare, NLE per salti piccoli, WAB per quelli medi e

DFS per salti grandi. Per quanto riguarda il controllo della varianza NLE risulta la carta migliore, ma l'efficienza di DFS è di praticamente allo stesso livello.

In conclusione, nonostante la carta WAB sia molto buona per identificare alcuni scostamenti dalla media per certe distribuzioni, la sua inefficacia nell'individuare cambiamenti della varianza, ci fa desistere dall'utilizzarla. Per quanto riguarda le carte NLE e DFS, sotto distribuzione normale, quest'ultima si rivela essere non solo la migliore carta non parametrica, ma anche meglio di CEW (per certi salti della media e per tutti i salti della varianza). Sotto distribuzione t di student, tranne che per un singolo caso (quando la media varia di δ = 0.25), DFS ha efficienza migliore di NLE. Sotto distribuzione χ^2 NLE e DFS sono in egual modo le più efficienti. In definitiva, che si tratti di variazioni della media o sregolazioni della varianza, la carta di controllo DFS si rivela essere molto efficiente e anche adeguatamente robusta rispetto alla concorrenza, sia essa di tipo parametrico o non parametrico.

CAPITOLO 4: Implementazione e applicazione di DFS su R

In questo capitolo andrò prima a mostrare e spiegare le funzioni definite in R che ho utilizzato per implementare DFS e poi passerò all'applicazione della carta a due diversi dataset: nel primo le osservazioni sono individuali, mentre nel secondo i dati sono raggruppati. I dati osservati nel primo dataset sono tutti in controllo, quindi per valutare le prestazioni di DFS, andrò a separare i dati in due parti: la prima mi sarà utile per stimare la situazione in controllo, mentre alla seconda parte andrò a modificare una parte dei dati, (da una certa osservazione in poi) in modo da simulare un aumento/diminuzione della varianza o della media, o di entrambe, al fine di constatare se la carta in esame segnala una situazione OC e quanto "tempo" ci impiega. Per quanto riguarda il secondo dataset invece, i dati risultano essere già suddivisi in queste due parti: dati IC di fase I e dati di fase II, che andrò a monitorare.

4.1) Codice R utilizzato per l'implementazione di DFS

In questa sezione andrò a presentare le funzioni necessarie ad implementare la carta DFS in R. Tra le varie funzioni si trova anche `ecdf_mod()`, una modifica della funzione di ripartizione empirica di R, questo perché nonostante gli Autori dell'articolo non trattino l'argomento, avere una funzione che stimi la funzione di ripartizione empirica in modo che essa abbia codominio in nell'insieme aperto (0,1) è fondamentale. Un'altra questione riguarda invece i limiti di controllo della carta DFS: mentre nell'articolo è stato presentato un caso in cui i limiti sono statici (una linea orizzontale), in questa implementazione si è optato per dei limiti dinamici calcolati via simulazione. Questa scelta è stata fatta sia per l'efficienza che ci garantiscono, sia per la relativa facilità con cui si calcolano.

- **`score(u)`:** riceve come argomento u, un oggetto che contiene i dati trasformati grazie alla funzione di ripartizione stimata grazie ai dati IC di Fase I. Questa funzione restituisce lo score di verosimiglianza, definito nell'articolo come: $\phi_1(\cdot)$ e $\phi_2(\cdot)$

```
score <- function(u) {
u <- as.matrix(u)
a <- 2 * u - 1
rbind(colSums(a), colSums(a * log(u / (1 - u)) - 1)) }
```

- **`ewma.score.crit()`:** questa funzione calcola i limiti dinamici L_t per simulazione. Riceve in ingresso una costante di lisciamento `lambda`, il numero di osservazione nei sottogruppi `n`, l'ARL in controllo desiderata `ARL`, l'istante di tempo t dal quale considerare i limiti L_t come costanti `tmax`, una costante utile per il lisciamento dei limiti posta pari a 50 `textra` e il numero di simulazioni su cui basare il calcolo `Nsim` (posto di default abbastanza elevato perché i limiti li calcoleremo solo una volta). La funzione restituisce una lista che contiene, oltre ai suoi argomenti in ingresso, anche i limiti dinamici L_t calcolati, sia in versione lisciata, sia in versione grezza.

```
ewma.score.crit <- function(lambda, n, ARL,
tmax = round(log(0.001/lambda)/log(1-lambda),digits = 0),
textra = 50, Nsim = max(10000, 20 * ARL)) {
    Lraw <- numeric(tmax + textra)
    Qstar <- 0
    Wstar <- rep(0, Nsim)
    NN <- n * Nsim
    sds <- sd.score * sqrt(n)
    for (i in 1:(tmax + textra)) {
          # calcolo degli Nsim score simulati
          score.sim <- score(matrix(runif(NN), n))
          # calcolo delle Nsim stat di controllo IC
          Qstar <- lambda * score.sim + (1 - lambda) * Qstar
          Wstar <- colSums((Qstar / sds)^2)
          # calcolo del quantile (1-1/B)
          Lt <- quantile(Wstar, 1 - 1 / ARL)
```

```
        # sostituzione degli Wstar OC
        idx <- which(Wstar > Lt)
        Qstar[,idx]  <-  Qstar[,sample(which(Wstar  <=  Lt),
length(idx))]
        Lraw[i] <- Lt
    }
    y <- -rev(isoreg(-rev(Lraw))$yf[-seq.int(textra)])
    x <- seq.int(tmax)
    I <- c(TRUE, y[-1] > y[-tmax])
    I[tmax] <- TRUE
    L <- approxfun(x[I], y[I])(x)
    invisible(list(
    lambda = lambda, n = n,
        ARL = ARL, tmax = tmax, textra = textra, Nsim = Nsim,
        Lraw = Lraw, # limiti grezzi (non lisciati)
        L = L # i limiti lisciati
   ))
}
```

- **`sd.score`:** questo oggetto contiene la diagonale della matrice di informazione di Fisher, in quanto con le assunzioni fatte nell'articolo, essa risulta una matrice diagonale che non dipende dai dati.

```
sd.score <- sqrt(c(1 / 3, (pi^2 + 3) / 9))
```

- **`ecdf_mod()`:** questa funzione, aggiungendo "mezza" osservazione a meno infinito (-∞) e mezza a più infinito (+∞), riporta il codominio della funzione di ripartizione empirica stimata all'intervallo aperto (0,1). Non utilizzare una funzione del genere creerebbe problemi nel momento in cui si va a calcolare lo score in 0 o 1.

```
ecdf_mod <- function(x) {
    n <- length(x)
    f <- ecdf(x)
    function(x) (0.5 + n * f(x)) / (n + 1)
}
```

- **`ewma.score()`:** questa funzione applica effettivamente la carta di controllo DFS ai dati, utilizzando i limiti calcolati in `ewma.score.crit()`; riceve poi un oggetto u, che contiene i dati trasformati grazie alla funzione di ripartizione stimata grazie alla funzione di ripartizione empirica calcolata in `ecdf_mod()` (con appunto un accorgimento per riportare il suo codominio in (0,1)). L'ultimo argomento della funzione, `plot` posto uguale a `TRUE`, permette la rappresentazione grafica della carta. La funzione restituisce poi una lista contenente gli istanti di tempo t ritenuti OC, i limiti dinamici, le due componenti che permettono di calcolare la statistica di controllo e infine la statistica di controllo, tutto questo per ogni istante di tempo t.

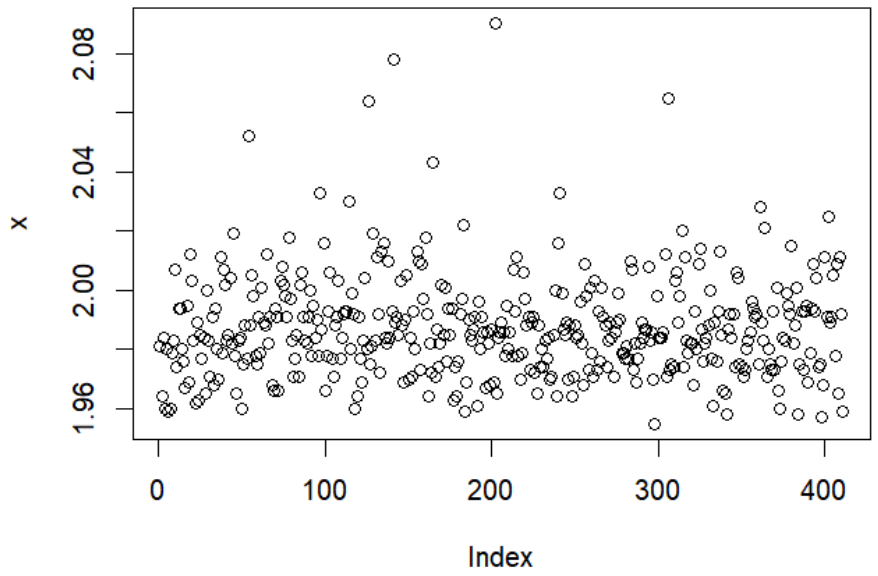
```
ewma.score <- function(chart, u, plot = TRUE) {
    # definizioni preliminari
    u <- as.matrix(u)
    m <- NROW(u)
    n <- NCOL(u)
    R <- numeric(m)
    L <- numeric(m)
    Q <- matrix(0, m, 2)

Qt <- numeric(2)
    sds <- sd.score * sqrt(n)
    for (i in 1:m) {
          # aggiornamento delle due EWMA
          Qt <- Qt + chart$lambda * (score(u[i,]) - Qt)
          ## Memorizziamo le due EWMA standardizzate
          Q[i, ] <- Qt / sds
          # calcolo della statistica di controllo
          R[i] <- sum(Q[i, ] * Q[i, ])
          # memorizzazione dei limiti
          L[i] <- chart$L[min(i, chart$tmax)]
    }
    OC <- which(R > L)
    if (plot == TRUE) {
          matplot(cbind(R, L),
```

```
        type = c("h", "s"), xlab = "campione",
        ylab = "statistica di controllo"
        )
        points(OC, R[OC], col = 2, pch = 20)
    }
    invisible(list(OC = OC, R = R, L = L, Q = Q)) }
```

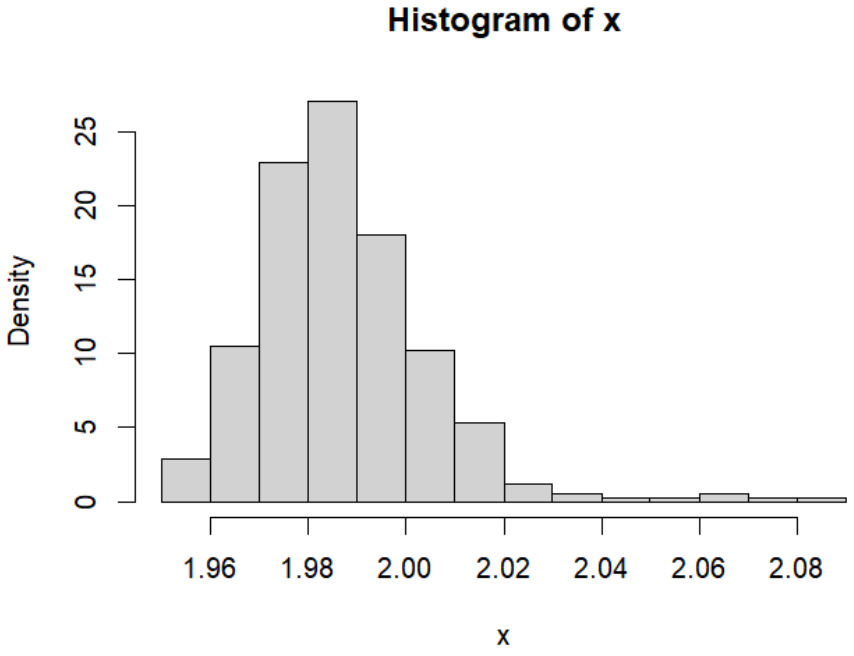
4.2) Applicazione della carta DFS ai dati del dataset “med.dat”

I dati raccolti in questo dataset forniscono il diametro (in mm) ad una delle due estremità di 411 tubicini di gomma che sono montati in una apparecchiatura medicale, e cambiati ad ogni paziente. Per poter essere fissato correttamente nell’apparecchiatura a cui è destinato, il diametro del tubicino deve essere compreso nell’intervallo 2 mm ± 0.1 mm, questi sono dunque i limiti di specifica. Procediamo con la lettura dei dati:

```
dati <- read.table("med.dat") # leggiamo il file med.dat
head(dati)
    V1
1 1.981
2 1.964
3 1.984
4 1.960
5 1.980
6 1.959
range(dati)
[1] 1.955 2.090
x <- dati$V1
plot(x) # diagramma di dispersione
```

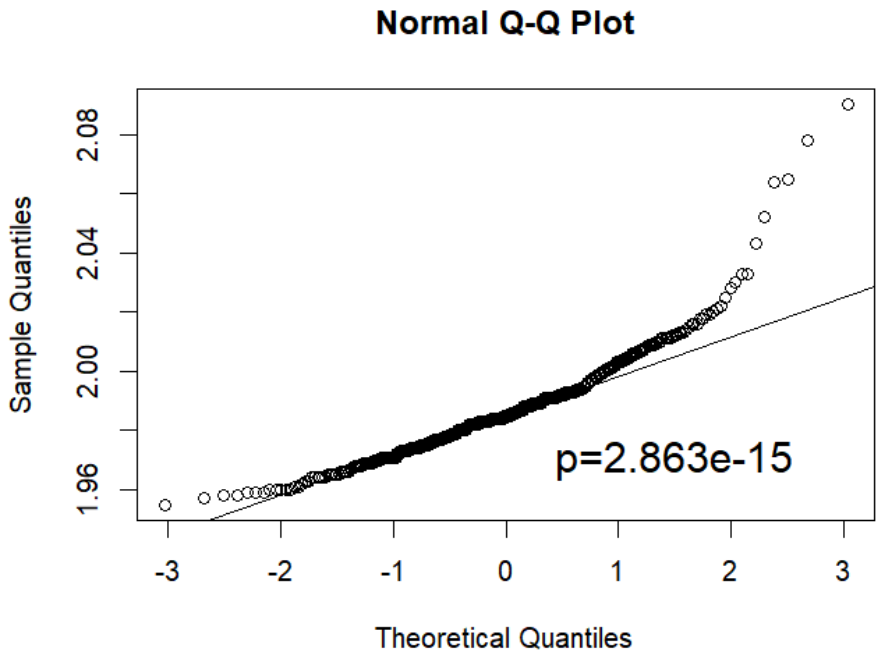



```
hist(x, prob = TRUE)
```



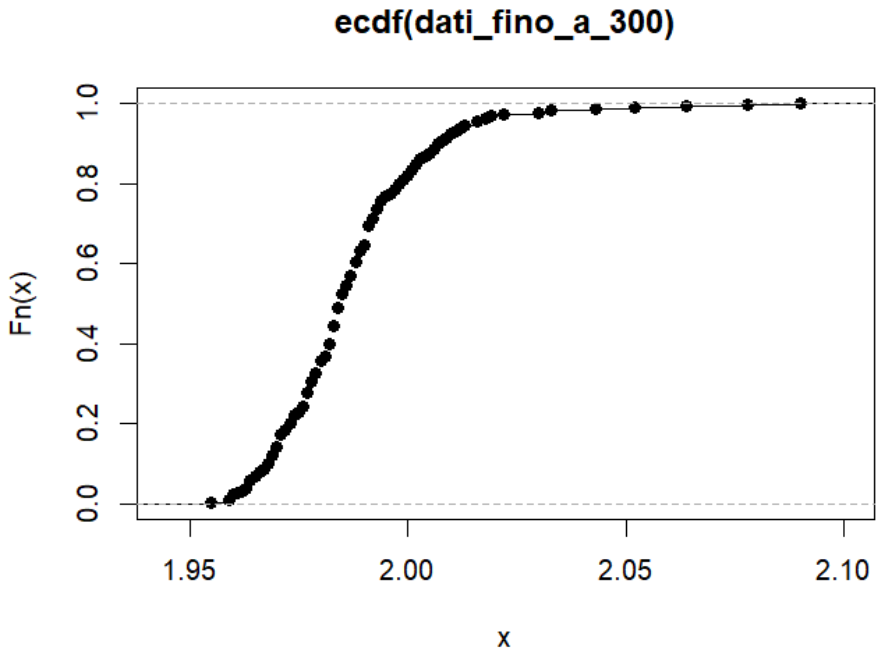
Si noti la forte asimmetria a destra dei dati, che un occhio attento può riconoscere anche dal diagramma di dispersione. Andiamo ora a vedere il grafico quantile-quantile e a fare un test sulla Normalità della distribuzione.

```
qqnorm(x) # grafico quantile-quantile
text(1.5, 1.97, paste("p=", round(shapiro.test(x)$p.value,
18), sep = ""), cex = 1.5) # test sulla Normalità
qqline(x) # retta teorica sulla quale si troverebbero i punti
# se la normalità fosse rispettata
```



Il test di Shapiro-Wilk per testare la Normalità e il grafico quantile-quantile, ci confermano la forte non normalità dei dati.

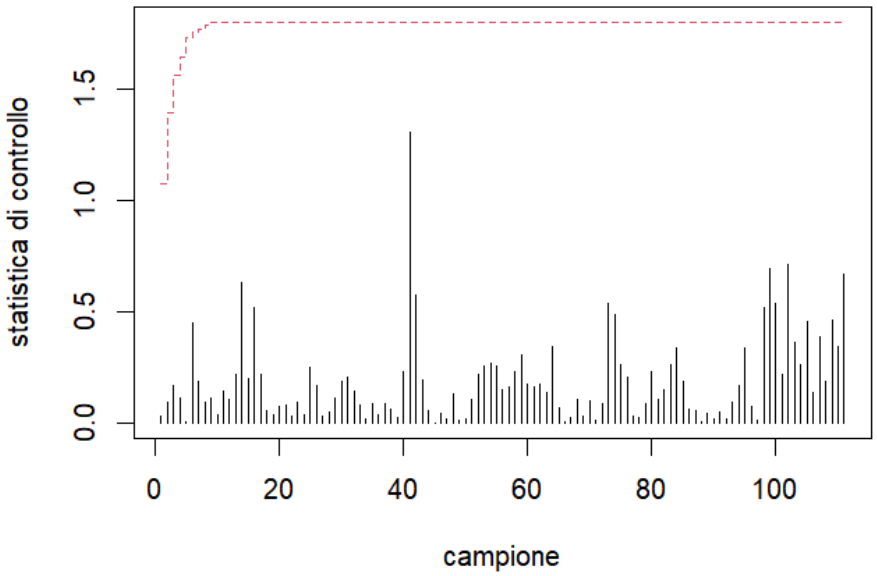
```
# definizione delle funzioni utili per implementare la carta
# DFS
dati_fino_a_300 <- dati[1:300,] # prendiamo la parte di dati
# utile per stimare la funzione di ripartizione (dati di Fase
# I, situazione in controllo IC)
mu0 <- mean(dati_fino_a_300) # media campionaria dei
# dati_fino_a_300
sd0 <- sd(dati_fino_a_300) # standard error dei
dati_fino_a_300
plot(ecdf(dati_fino_a_300)) # disegniamo la funzione di
# ripartizione empirica stimata durante l'ipotetica Fase I
```



Abbiamo appena definito la parte di dati coi quali stimare la funzione di ripartizione empirica, e l'abbiamo disegnata; bisogna ricordarsi che questa funzione, se non modificata per riportare il suo codominio in (0,1), creerebbe problemi nel calcolo dello score. Andiamo quindi ad applicare la nostra funzione `ecdf_mod()` ai dati di fase I e successivamente a calcolare i limiti dinamici e ad applicare la carta DFS ai `dati_dopo_300`.

```
Funzione_F <- ecdf_mod(dati_fino_a_300) # funzione di
# ripartizione empirica
dati_dopo_300 <- dati[301:411, ] # dati che vogliamo
monitorare
u <- Funzione_F(dati_dopo_300) # applichiamo la funzione di
# ripartizione empirica ai dati che vogliamo monitorare,
ovvero # ai dati_dopo_300
range(u) # adesso sì che u non potrà mai assumere valori 0 o 1
[1] 0.004983389 0.991694352
# calcolo dei limiti dinamici
chart <- ewma.score.crit(0.2, 1, 500, Nsim = 500000)
# calcoliamo i limiti dinamici utilizzando una costante di
# lisciamento pari a 0.2, osservazioni individuali, ARL IC
```

```
# desiderato pari a 500 e un numero di simulazioni molto
# elevato, applichiamo la carta ai dati da monitorare (quelli
dopo 300)
a <- ewma.score(chart = chart, u = Funzione_F(dati_dopo_300),
plot = TRUE)
a$OC
integer(0)
```



Come possiamo vedere, la carta non chiama mai un allarme, questo era da aspettarselo in quanto in realtà tutti i dati erano IC; inoltre l'ARL IC scelto è abbastanza grande rispetto alla dimensione del campione (500 vs 111), per cui non ci aspettiamo falsi allarmi. Nonostante la carta non segnali alcun allarme, si possono fare comunque certe considerazioni: la statistica di controllo ha un valore abbastanza elevato per l'osservazione 41, andiamo ad indagare.

```
which.max(a$R)
[1] 41
max(a$R)
[1] 1.303268
```

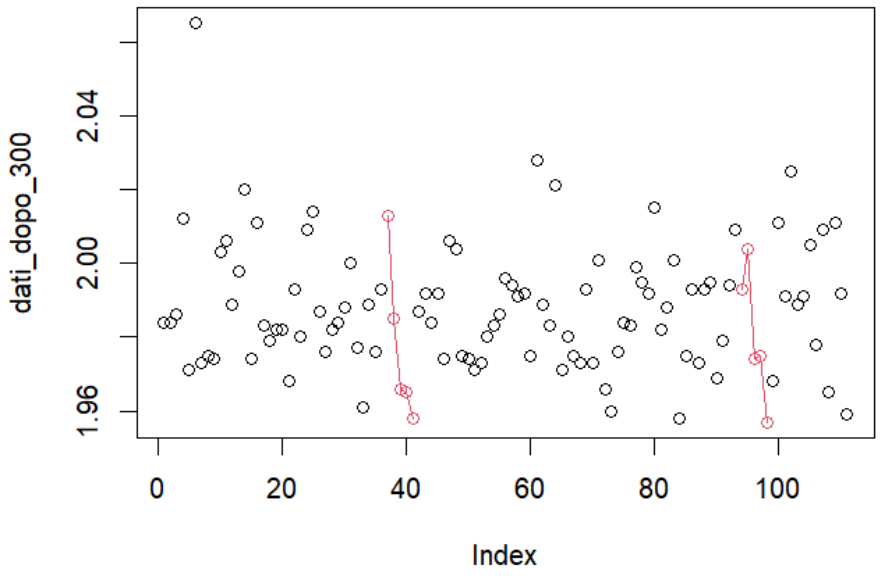
Si può notare come l'osservazione 41 sia la seconda osservazione più piccola dei `dati_dopo_300` e inoltre il valore delle due osservazioni è molto vicino tra loro:

```
which.min(dati_dopo_300)
[1] 98
which.min(dati_dopo_300[-which.min(dati_dopo_300)])
[1] 41
dati_dopo_300[c(98,41)]
[1] 1.957 1.958
```

Però non conta solo il valore dell'osservazione al tempo t al fine di chiamare o meno un allarme, contano anche le osservazioni ai tempi precedenti (oltre che alla forma dello score), e dunque andiamo a vedere le osservazioni precedenti:

```
plot(dati_dopo_300)
points(c(37:41),dati_dopo_300[c(37:41)],col=2,type = "o")
points(c(94:98),dati_dopo_300[c(94:98)],col=2,type = "o")
```

Si vede come le misure precedenti alla 41 siano strettamente decrescenti, mentre non si può dire lo stesso di quelle precedenti alla 98.

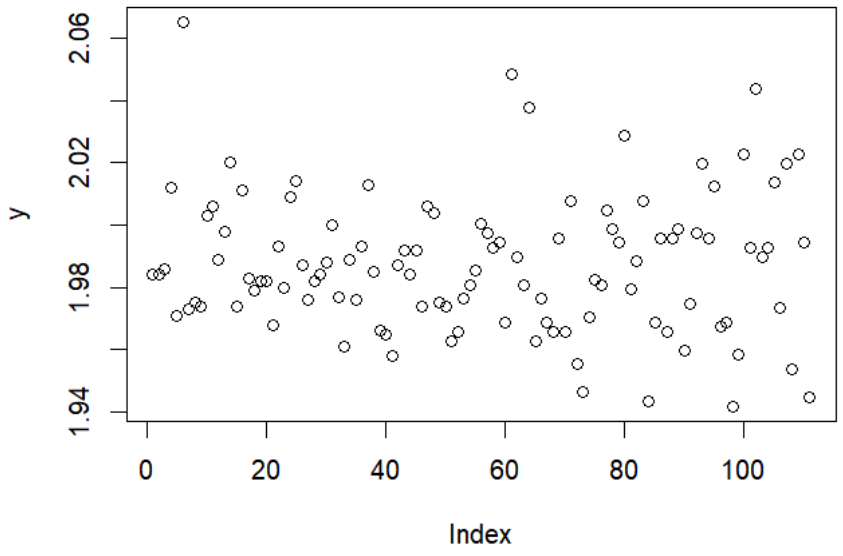


Al fine di poter “toccare con mano” la capacità della carta DFS di segnalare allarmi, andrò ora a presentare dei casi in cui parte dei dati analizzati sono stati alterati. Visto che gli Autori dell’articolo non ne hanno parlato, ho cercato un criterio per capire, al momento del segnale, che cosa è cambiato. In teoria non è complicato, visto che abbiamo uno score per i cambiamenti di posizione e uno per i cambiamenti di dispersione; il criterio che ho utilizzato consiste semplicemente nello standardizzare le due componenti dello score dell’istante di tempo in cui è stato chiamato il primo allarme, in modo che la loro somma sia 1. In questo modo per individuare quale delle due componenti della carta è più responsabile per la chiamata dell’allarme, basterà vedere quale componente è maggiore di 0.5 (oppure se sono entrambe vicine a 0.5 allora è possibile che il problema da ricercare riguardi entrambe).

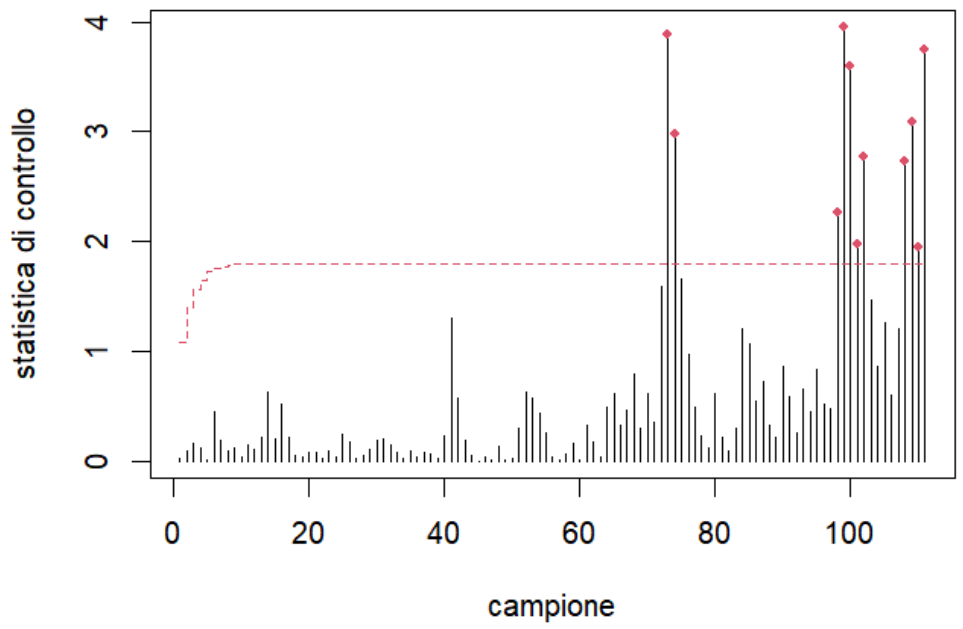
4.2.1) Caso 1: dati con un aumento (solo) della varianza

Dati con un aumento in varianza dopo l'osservazione 351, compresa. Andiamo ad aumentare del 50% la varianza dei dati, poi applichiamo la carta agli stessi:

```
y <- dati_dopo_300
y[51:111] <- mu0 + 1.5 * (y[51:111] - mu0)
plot(y)
```



```
b <- ewma.score(chart, Funzione_F(y), plot = TRUE)# applico la
# carta
b$OC # istanti fuori controllo
[1]  73  74  98  99 100 101 102 108 109 110 111
```



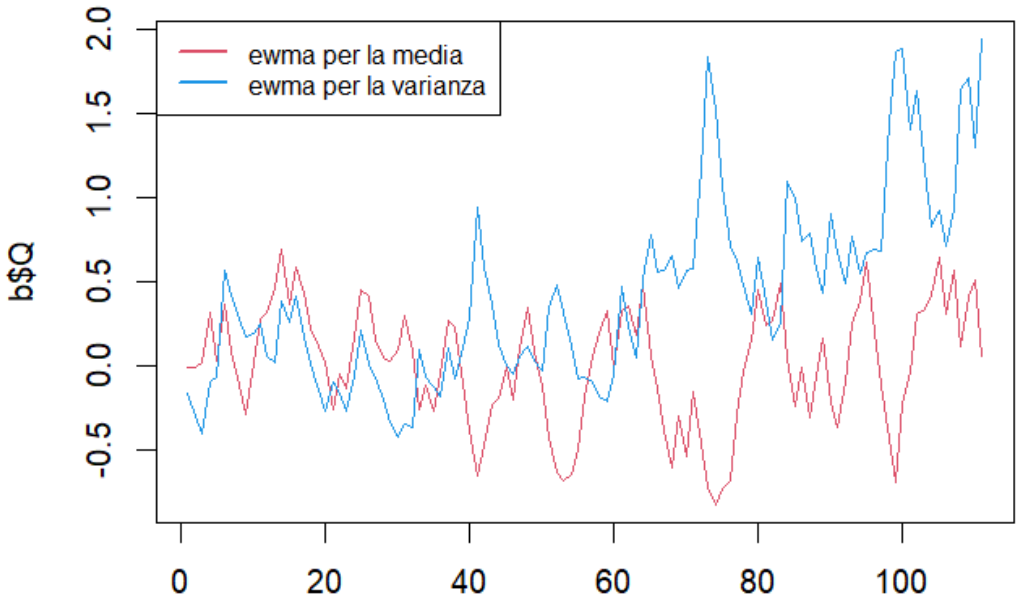
Per accorgersi di un aumento della varianza del 50% (quindi abbastanza grande), sono necessarie:

```
min(b$OC) - 51
[1] 22
```

22 osservazioni, che non sono poche, ma bisogna tenere conto che stiamo usando una carta non parametrica; inoltre individuare scostamenti dalla situazione in controllo per la varianza, è “più complicato” che per la media, soprattutto nel caso di osservazioni individuali, come nel caso in esame.

Vediamo ora un grafico che rappresenta come fluttuano le due parti dello score al passare dei campioni; in rosso è colorata l'ewma per la media, mentre in blu l'ewma per la varianza, esse sono standardizzate in `ewma.score()` quindi sono direttamente confrontabili.

```
matplot(b$Q, type = "l", lty = 1, col = c(2, 4)) #il grafico
legend("topleft",legend = c("ewma per la media", "ewma per la
varianza"), col = c(2, 4), lwd = 2, cex=.8) #la legenda
```



Si vede come dall'osservazione 64 in poi, il contributo alla statistica di controllo che "monitora" la varianza, è sempre maggiore dell'altra componente che "monitora" la media, almeno fino al picco raggiunto all'osservazione 73. Proprio a questa osservazione, la carta chiama il primo allarme; ricordandoci che stiamo utilizzando una carta di fase II, il resto del grafico risulta non molto significativo, in quanto in una situazione reale, avremmo interrotto la produzione e smesso di raccogliere dati.

```
b$Q[73,]
[1] -0.7163024  1.8380127
# contributo delle due EWMA alla statistica di controllo nel
# momento del primo allarme
tm <- min(b$OC)
tm # momento del primo allarme
[1] 73
b$Q[tm, ]^2 / b$R[tm]
```

```
[1] 0.1318527 0.8681473
```

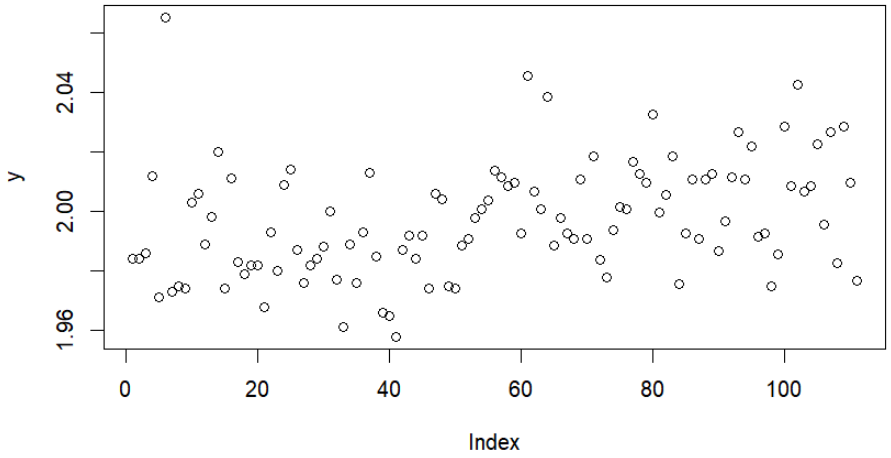
Da qui si capisce che la componente che monitora la varianza, al momento dell'allarme, *spiega* più dell'86% della statistica di controllo, dunque questo banale calcolo ci permette di dire che: dato che la carta ha chiamato un allarme, questo "probabilmente" è dovuto a problemi in varianza, piuttosto che in media.

4.2.2) Caso 2: dati con un aumento (solo) della media

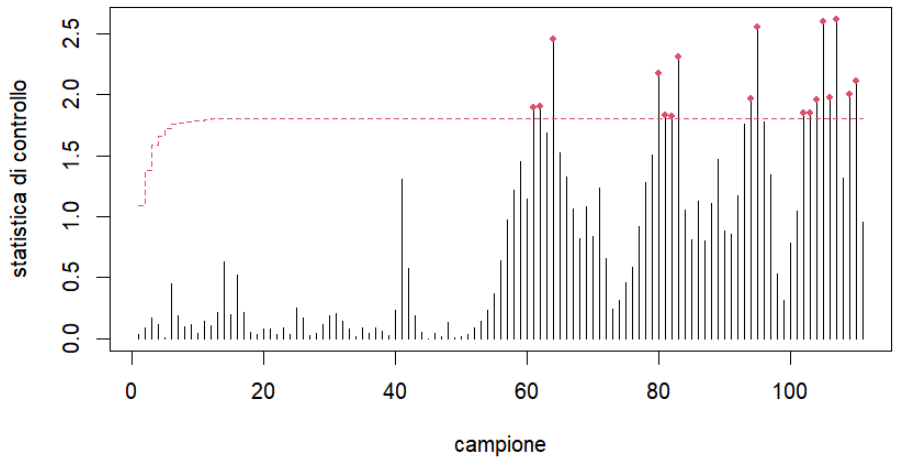
Dati con un aumento in media dopo l'osservazione 351, compresa. Andiamo ad aumentare la media dei dati di una volta lo standard error campionario

```
y <- dati_dopo_300
y[51:111] <- y[51:111] + 1 * sd0
plot(y)
```

Si vede come la seconda parte dei dati sia in media più grande rispetto alla prima




```
b <- ewma.score(chart, Funzione_F(y), plot = TRUE)# applico la carta
b$OC # istanti fuori controllo
[1]  61  62  64  80  81  82  83  94  95 102 103 104 105 106 107 109 110
```

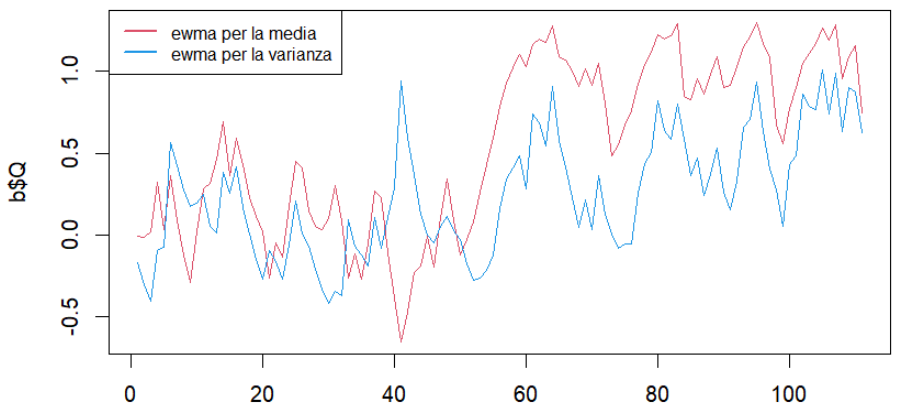



```
min(b$OC) - 51
[1] 10
```

La carta è piuttosto efficiente e impiega solo 10 istanti di tempo prima di chiamare l'allarme. Come commentato in precedenza, individuare scostamenti in media, anche piuttosto piccoli, (come in questo caso) è più semplice rispetto ad individuare quelli in varianza, soprattutto nel caso di osservazioni individuali (situazione in cui ci troviamo).

Rappresento ora il grafico dell'andamento delle due componenti dello score al passare delle osservazioni

```
matplot(b$Q, type = "l", lty = 1, col = c(2, 4)) # il grafico
legend("topleft",legend = c("ewma per la media", "ewma per la
varianza"), col = c(2, 4), lwd = 2, cex=.8) # la legenda
```



Andiamo ora a vedere il contributo delle due EWMA standardizzate, alla statistica di controllo nel momento del primo allarme:

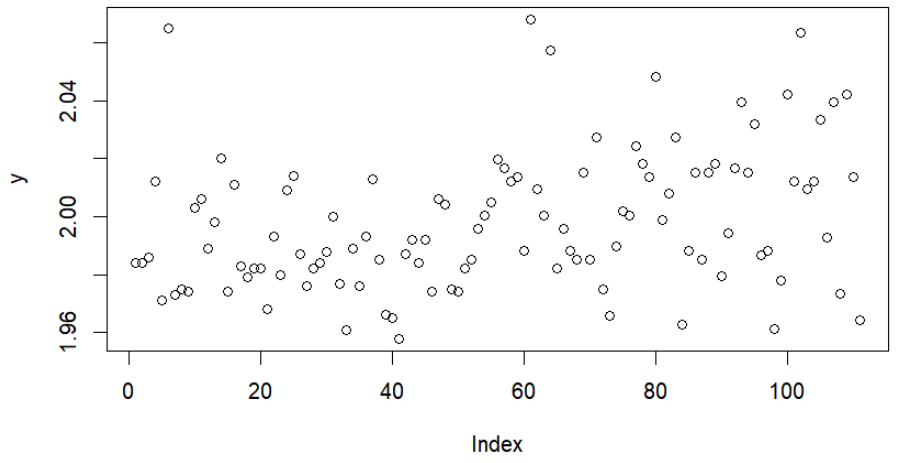
```
tm <- min(b$OC)
tm
[1] 61
b$Q[tm, ]^2 / b$R[tm]
[1] 0.7110102 0.2889898
```

Da quest'ultimo calcolo si capisce come la componente dell'ewma che monitora la media, al momento in cui la carta chiama il primo allarme, *spiega* più del 70% della statistica di controllo. Concludiamo dicendo che al fine di riportare la situazione in controllo, bisognerà "probabilmente" indagare cosa ha modificato la media della distribuzione (piuttosto che la varianza).

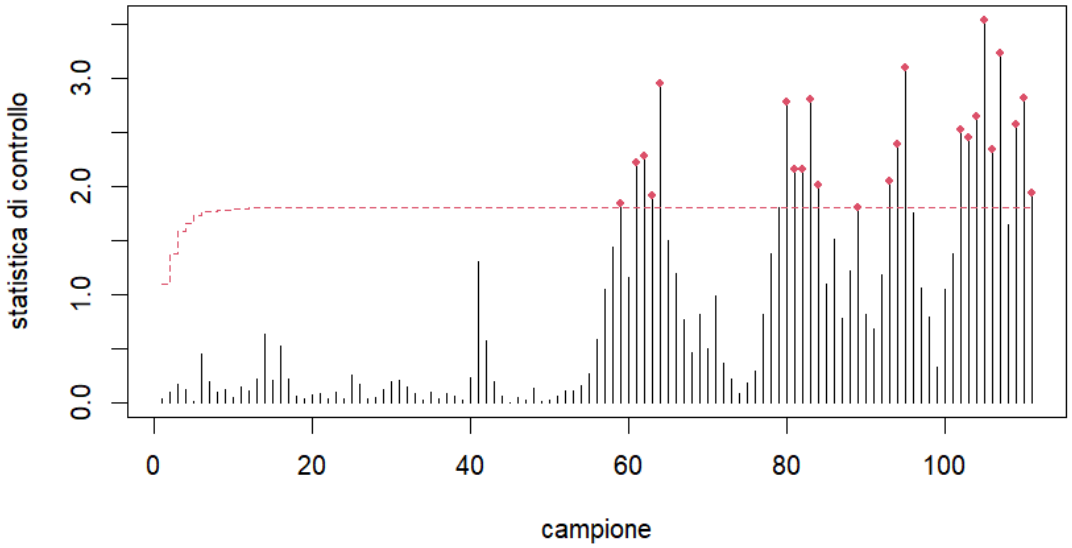
4.2.3) Caso 3: dati con un aumento sia della media che della varianza

Dati con un aumento in media e in varianza dopo l'osservazione 351, compresa. Andiamo ad aumentare la media di 1.1 volte lo standard error campionario e di 1.5 volte la varianza dei dati considerati in controllo:

```
y <- dati_dopo_300
y[51:111] <- mu0 + 1.1*sd0 + 1.5*(y[51:111] - mu0)
plot(y) #da questo grafico si vede proprio come la seconda
# parte dei dati abbia una media un po' più elevata e anche
(si # nota più chiaramente), una varianza più grande
```



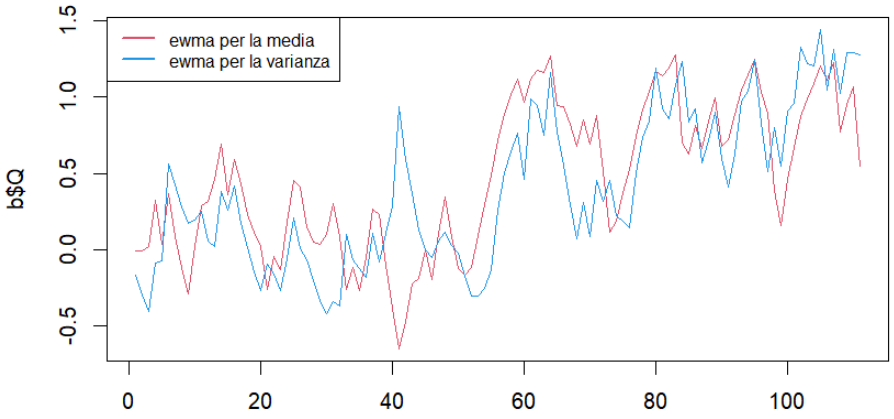
```
b <- ewma.score(chart, Funzione_F(y), plot = TRUE)
# applichiamo la carta
b$OC #istanti fuori controllo
[1]  59  61  62  63  64  80  81  82  83  84  89  93  94  95 102 103 104
105 106 107 109 110 111
```




```
min(b$OC) - 51
[1]  8
```

In questa situazione la carta è piuttosto efficiente, quando cambiano sia la media che la varianza, la carta individua ancora prima il passaggio del processo ad uno stato fuori controllo OC (rispetto al caso in cui aumentano solo delle due caratteristiche della distribuzione) e chiama l'allarme dopo sole 8 osservazioni.
Rappresento ora il grafico dell'andamento delle due componenti dello score al passare delle osservazioni

```
matplot(b$Q, type = "l", lty = 1, col = c(2, 4)) # il grafico
legend("topleft",legend = c("ewma per la media", "ewma per la
varianza"), col = c(2, 4), lwd = 2, cex=.8) # la legenda
```




```
# andiamo ora a vedere il contributo delle due EWMA
# standardizzate alla statistica di controllo nel momento del
# primo allarme
tm <- min(b$OC)
tm
[1] 59
b$Q[tm, ]^2 / b$R[tm]
[1] 0.6805153 0.3194847
```

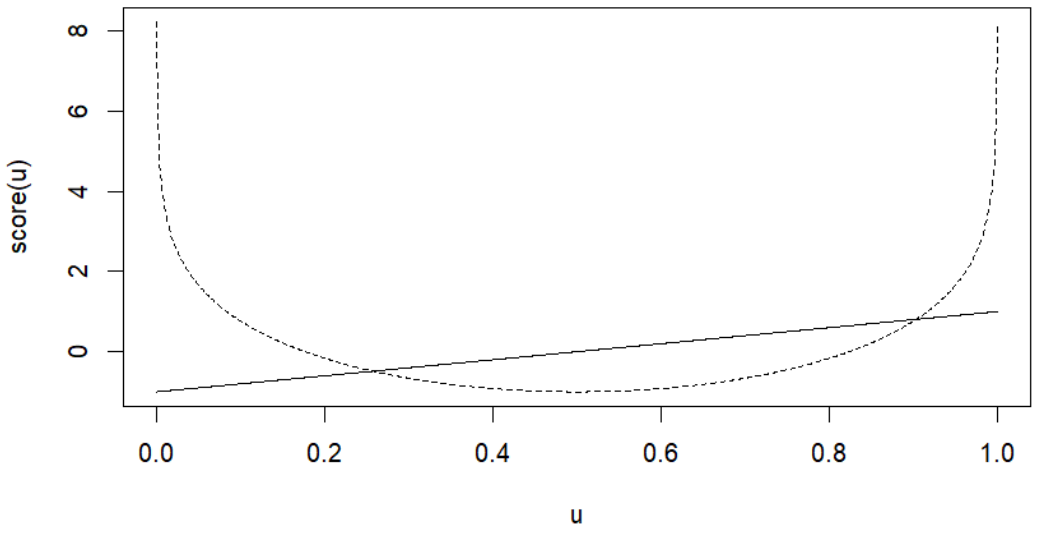
In questo caso vediamo che, seguendo il nostro criterio di valutazione, saremmo più inclini a dire che il fuori controllo è stato chiamato per problemi che riguardano la media (della dimensione dei tubicini). Purtroppo però il nostro criterio non riesce ad identificare che la vera causa speciale di variabilità ha aumentato entrambe. Bisogna però tenere in considerazione che identificare scostamenti in media è più "facile" per la nostra carta di controllo (rispetto quelli in varianza), per questo motivo il nostro criterio ci fornisce questa risposta.

4.2.4) Riflessione riguardo la forma dello score utilizzato dalla carta DFS

A questo punto penso che una riflessione riguardo la forma (scelta) dello score della carta DFS, sia doveroso farla, andiamo dunque a disegnare lo score:

```
u <- matrix(seq(1E-04, 1-1E-04, length=1000), 1)
s <- score(u)
matplot(as.numeric(u), t(s), type="l", col=1, lty=1:2,
        xlab="u", ylab="score(u)")
```

Dal grafico seguente possiamo capire che: se la distribuzione delle u si sposta verso sinistra o verso destra, cosa che capita quando ho un cambiamento in media, allora anche lo score per la varianza (per come è definito) aumenta, di conseguenza il nostro criterio che cerca di farci capire se l'allarme che è stato chiamato dipende da problemi in media o in varianza, sarà sempre un po' starato, in quando lo score per la media è legato a quello per la varianza (la formula dello score per la media è proprio contenuta anche nello score per la varianza).



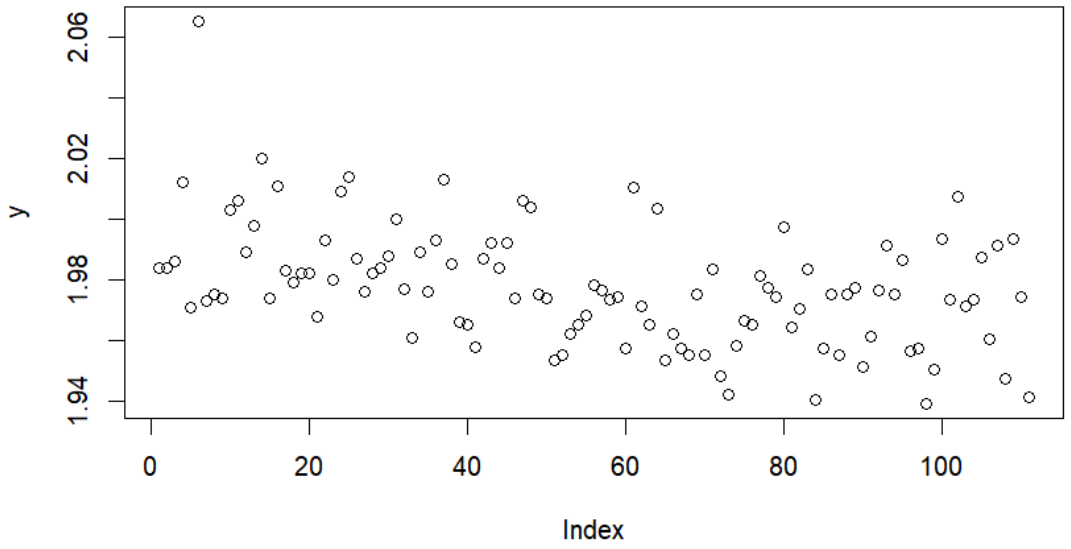
Un'altra cosa da aggiungere riguardo la forma dello score (e lo capiremo bene nel caso 5), riguarda il fatto che proprio per il fatto che il nostro score per la varianza ha forma ad "U", se u (variabile nelle ascisse del grafico sopra) si avvicina a 0 o ad 1, lo score di u esplode, ovvero diventa molto grande rapidamente e ciò permette a DFS di chiamare velocemente un allarme; d'altra parte se u è vicino a 0.5, lo score diminuisce troppo poco rapidamente e non ci permetterà di chiamare un allarme nel caso di diminuzione della varianza, ma quest'ultima affermazione è valida solo per il caso di osservazioni individuali,

mentre per dati raggruppati, la carta dovrebbe essere in grado di segnalare allarmi anche per la diminuzione della varianza

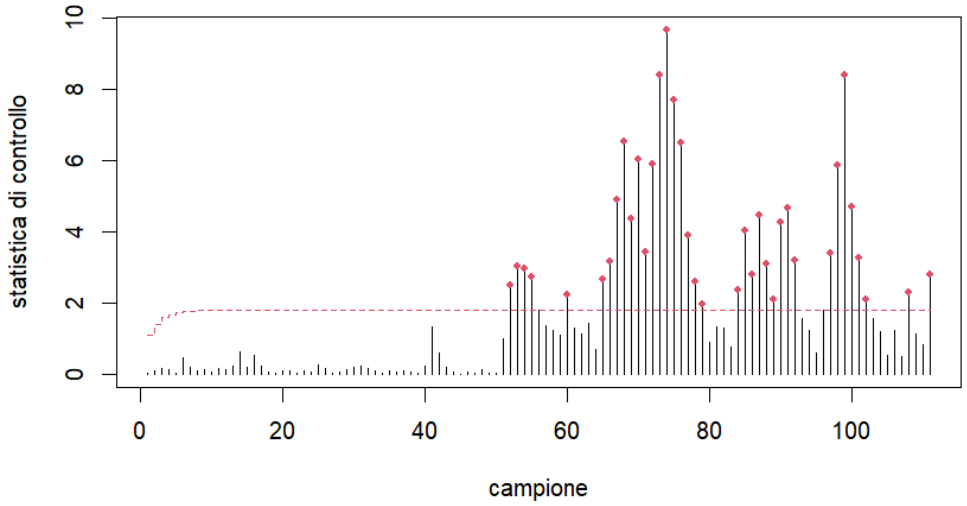
4.2.5) Caso 4: dati con una diminuzione (solo) della media

Dati con una diminuzione in media dopo l'osservazione 351, compresa. Andiamo a diminuire la media di una volta lo standard error campionario.

```
y <- dati_dopo_300
y[51:111] <- y[51:111] - 1 * sd0
plot(y) #si vede come la seconda parte dei dati sia in media
# più piccola della prima
```




```
b <- ewma.score(chart, Funzione_F(y), plot = TRUE)
#applichiamo # la carta
b$OC
[1]  52  53  54  55  60  65  66  67  68  69  70  71  72  73  74  75  76
77  78  79  84  85  86  87  88  89  90  91  92  97  98  99 100 101 102
108 111
```

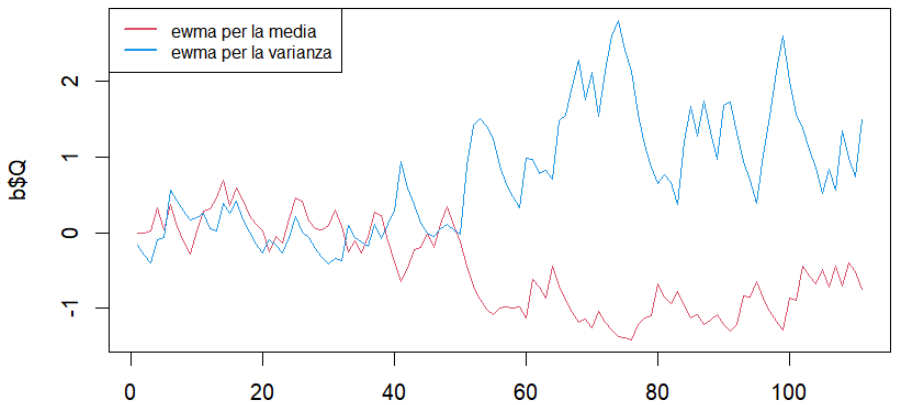



```
min(b$OC) - 51
[1] 1
```

La carta è super efficiente ed impiega solo 1 istante di tempo prima di chiamare l'allarme, questo perché come abbiamo visto dalla distribuzione dei dati, essa è asimmetrica ed abbiamo molta più informazione riguardo dove "finisca" la coda sinistra (piuttosto che quella destra), della vera distribuzione (ignota) da cui pensiamo possano provenire i dati osservati.

Rappresento ora il grafico dell'andamento delle due componenti dello score al passare delle osservazioni:

```
matplot(b$Q, type = "l", lty = 1, col = c(2, 4))
legend("topleft",legend = c("ewma per la media", "ewma per la
varianza"), col = c(2, 4), lwd = 2, cex=.8)
```



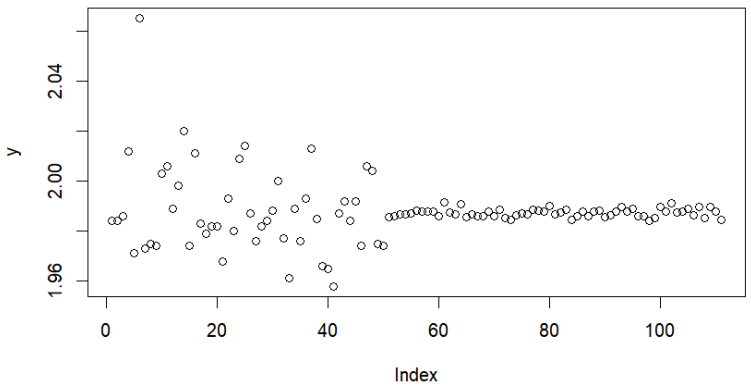
```
# andiamo ora a vedere il contributo delle due EWMA
# standardizzate alla statistica di controllo nel momento del
# primo allarme
tm <- min(b$OC)
tm
[1] 52
b$Q[tm, ]^2 / b$R[tm]
[1] 0.1961151 0.8038849
```

Quest'ultimo risultato e il grafico appena disegnato, porterebbero a conclusioni diametralmente opposte a quelle corrette: secondo il nostro criterio la carta sta chiamando un errore a causa della parte ewma che monitora la varianza, mentre ciò che è realmente accaduto è stata una diminuzione solo della media. La “colpa” di questo esito, come detto nella sezione 4.2.4), è della formulazione dello score di DFS: quando la parte dello score relativa all’ewma della media diminuisce, la relativa all’ewma della varianza aumenta, e per questo motivo il nostro criterio di valutazione fallisce nel suo obiettivo.

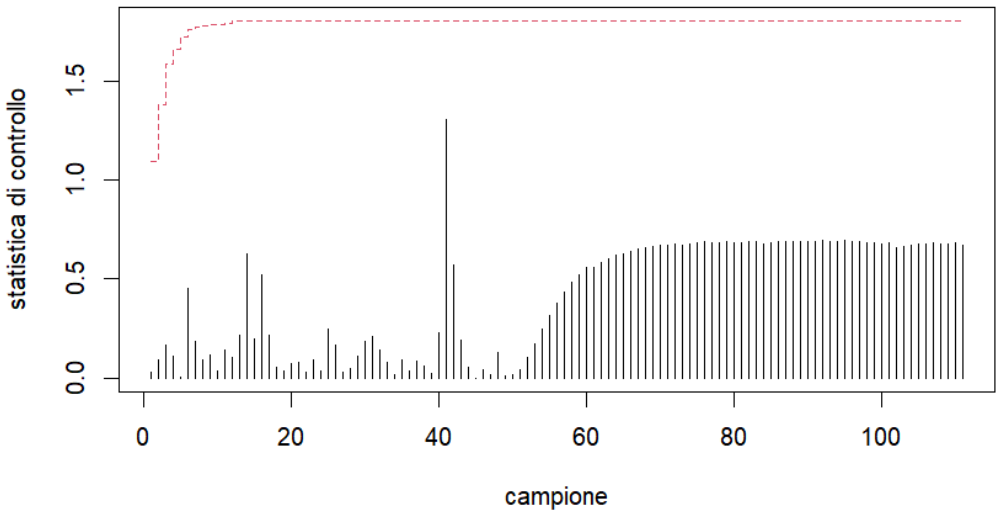
4.2.6) Caso 5: dati con una diminuzione (solo) della varianza

Dati con una diminuzione in varianza dopo l'osservazione 351 compresa, Andiamo a diminuire del 90% la varianza dei dati, si tratta volutamente di una situazione estrema, per far vedere una debolezza della carta DFS.

```
y <- dati_dopo_300
y[51:111] <- mu0 + .1 * (y[51:111] - mu0)
plot(y) #vediamo come i dati della seconda parte abbiano una
# varianza di molto inferiore a quelli della prima parte
```



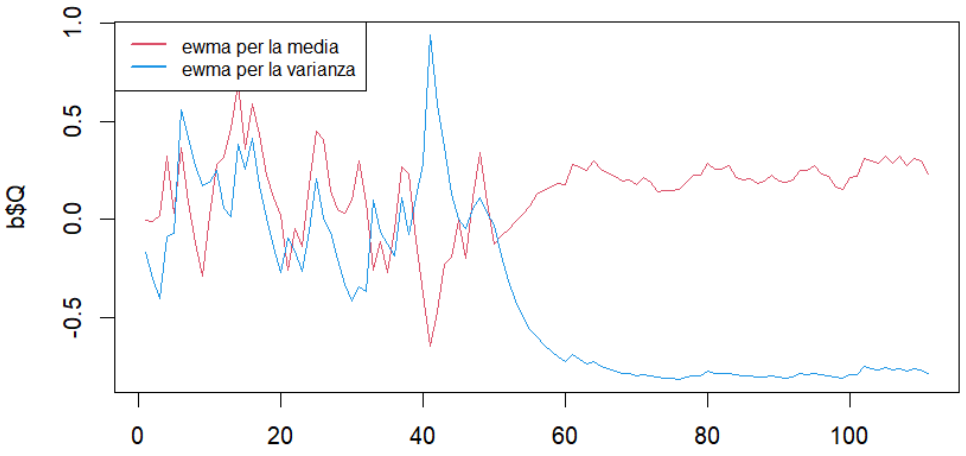
```
b <- ewma.score(chart, Funzione_F(y), plot = TRUE)
b$OC
integer(0)
```



Come anticipato nella sezione 4.2.4), anche se la varianza diminuisce di molto, la carta DFS non riesce a segnalare un allarme sia perché abbiamo osservazioni individuali, sia per la forma dello score: esso decresce troppo poco rapidamente quando le u sono vicino a 0.5.

Andiamo a vedere ugualmente il grafico dello score:

```
matplot(b$Q, type = "l", lty = 1, col = c(2, 4))
legend("topleft",legend = c("ewma per la media", "ewma per la
varianza"), col = c(2, 4), lwd = 2, cex=.8)
```



Da questo grafico possiamo capire una cosa: dopo che il processo è andato fuori controllo, le due componenti dello score si stabilizzano a certi livelli, e la carta non riesce a segnalare un allarme.
In conclusione c'è da dire che: non segnalare allarmi quando la varianza diminuisce, dal punto di vista operativo può anche essere utile, basti ricordare che la qualità di un processo è definita come l'inverso della sua variabilità in SPC; e dunque se per qualche motivo la varianza del processo si è abbassata, meglio per noi, vorrà dire che siamo stati fortunati e il processo è riuscito ad essere più capace di quanto ci aspettavamo e non è necessario chiamare un allarme.

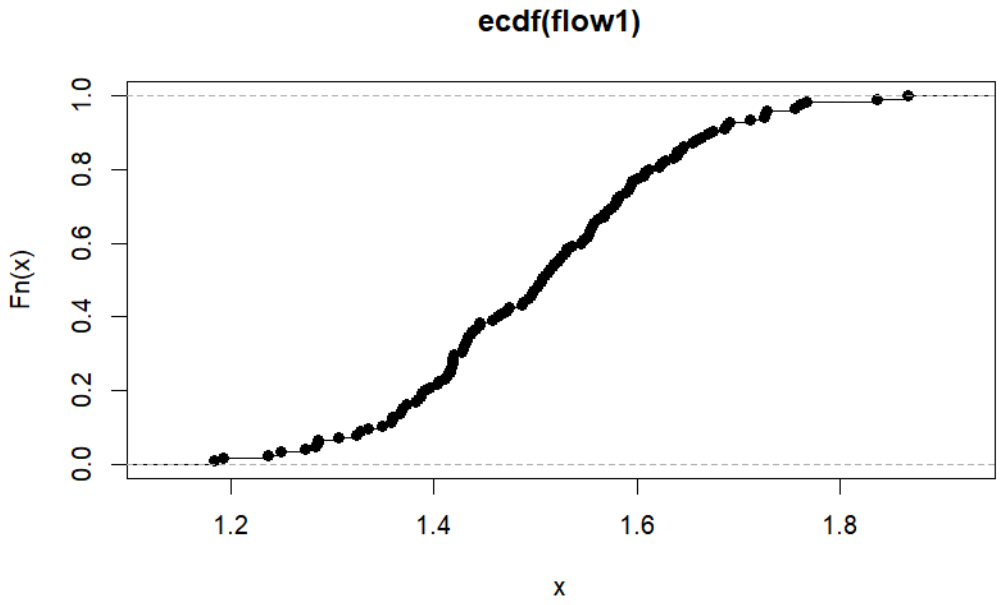
4.3) Applicazione della carta DFS ai dati del dataset "flow.dat"

I dati di questo dataset sono stati raccolti in un'industria che produce semiconduttori e riguardano l'espansione, (misurata in micron, ovvero, milionesimi di metro) dovuta alla cottura, di una vernice che viene spalmata sui wafer di silicio e poi appunto cotta al forno ad alte temperature.
Per ogni infornata, l'espansione è stata misurata su 5 wafer. La prima parte del dataset (fino all'osservazione 125), contiene le misure per 25 "infornate" consecutive, raccolte inizialmente per caratterizzarne e stimarne la distribuzione in controllo. Usando queste informazioni l'azienda ha poi costruito una carta di controllo Shewhart $\bar{X}$ per sorvegliare sequenzialmente la variabile considerata. I dati raccolti durante una sequenza di 20 ulteriori "infornate" (sempre 5 misure per ogni "infornata") sono contenuti nella parte restante del dataset.
Procediamo con la lettura dei dati:

```
# Un secondo caso studio:
#  proviamo  ad  applicare  la  carta  DFS  a  flow.dat  andiamo  a
# leggere il file che contiene i dati:
flow = scan("flow.dat")  #leggiamo il file flow.dat
```

Seguendo la descrizione del dataset, lo partiziono in due parti: dati di fase I e dati di fase II, questi ultimi li monitorerò con la carta DFS. Disegniamo poi la funzione di ripartizione empirica per i dati di fase I.

```
flow1 = flow[1:125] contiene i dati di fase I
flow2 =  flow[126:225] contiene i dati di fase II
n=5 # numerosità dei sottogruppi
m=25 # numero di sottogruppi
plot(ecdf(flow1))  #  disegniamo  la  funzione  di  ripartizione
# empirica stimata con i dati di Fase I
```

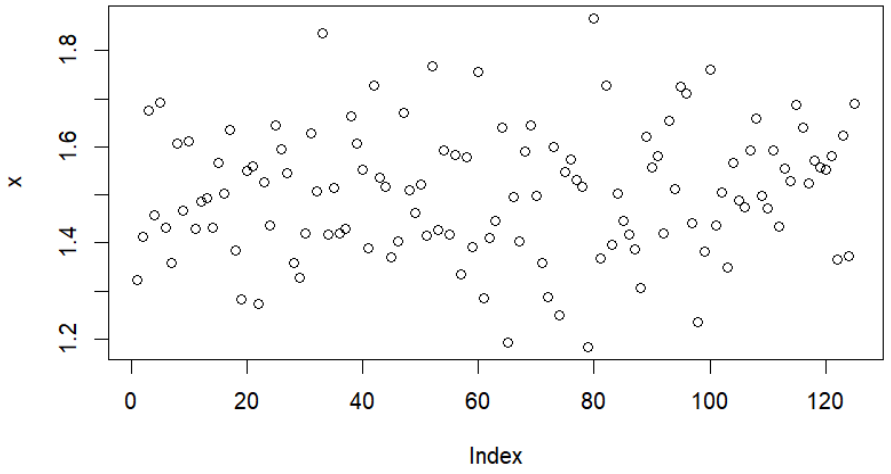


Come per il primo dataset, andiamo a calcolare la funzione di ripartizione empirica stimata con i dati di fase I e modificata adeguatamente per poterla poi applicare ai nostri dati

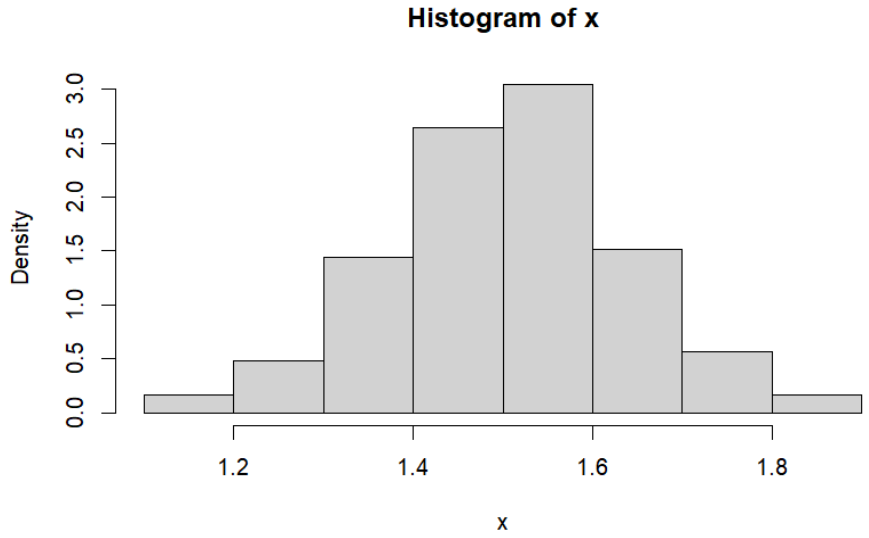
```
Funzione_F  <-  ecdf_mod(flow1)  #  funzione  di  ripartizione
empirica modificata per avere codominio (0,1)
u  <-  Funzione_F(flow2)  #  applichiamola  ai  dati  che  vogliamo
# monitorare
range(u)
[1] 0.003968254 0.996031746
# vediamo come nessuna u vale 0 o 1
```

Disegnamo ora il grafico di dispersione e l'istogramma dei dati:

```
x <- flow1
plot(x) # diagramma di dispersione dei dati di fase I
```

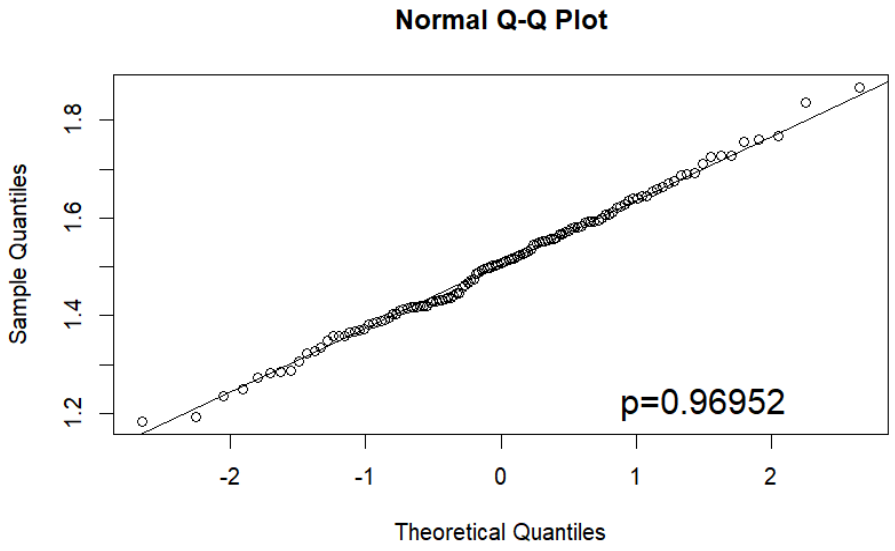



```
hist(x, prob = TRUE) # istogramma dei dati
```



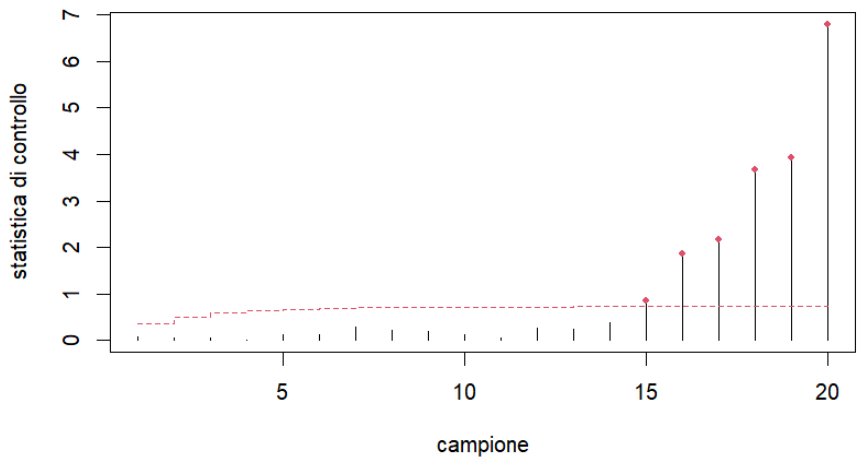
Sembra che i dati possano seguire una distribuzione normale, andiamo a verificarlo con un grafico quantile-quantile ed un test sulla Normalità.

```
qqnorm(x) # grafico quantile-quantile
qqline(x) # retta dei quantili teorici di una Normale
text(x=1.5,y=1.22
,paste("p=",round(shapiro.test(x)$p.value,5), sep = ""), cex =
1.5)
```



Accettiamo ampiamente l'ipotesi di Normalità dei dati. Procediamo ora ad applicare la carta DFS ai nostri dati, ma prima calcoliamoci i limiti dinamici e prepariamoci i dati in u, per poterla applicare:

```
chart <- ewma.score.crit(lambda = 0.2, n = n , ARL = 50, Nsim
= 500000) # calcoliamo i limiti
# prepariamoci u come una matrice (m x n)
u <- matrix(Funzione_F(flow2),nrow=chart$n)
u = t(u)
dim(u) #si tratta di una matrice 20x5
[1] 20  5
a <- ewma.score(chart = chart, u = u, plot = TRUE) #
calcoliamo # e disegniamo la carta di controllo
```



```
min(a$OC)
[1] 15
```

La carta DFS chiama l'allarme al 15esimo sottogruppo, che è molto prima del 20esimo, ovvero quando la carta di Shewhart $\bar{X}$ segnala l'allarme. La carta Shewhart $\bar{X}$ è quella che usava l'azienda per controllare il processo produttivo.

Utilizziamo ora il nostro criterio di valutazione:

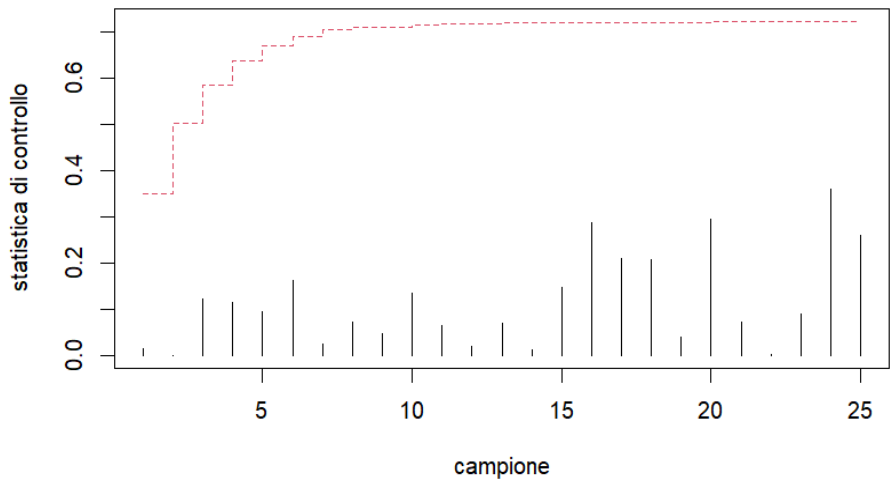
```
tm <- min(a$OC)
tm # istante di tempo del primo allarme
[1] 15
a$Q[tm, ]^2 / a$R[tm]
[1] 0.94109077 0.05890923
```

La carta, al momento dell'allarme, sta segnalando una situazione OC, segnalata principalmente (>94%) dalla componente dello score che monitora la media, quindi potremmo consigliare all'azienda di andare a controllare cosa potrebbe aver modificato la media dell'espansione della vernice sui wafer.

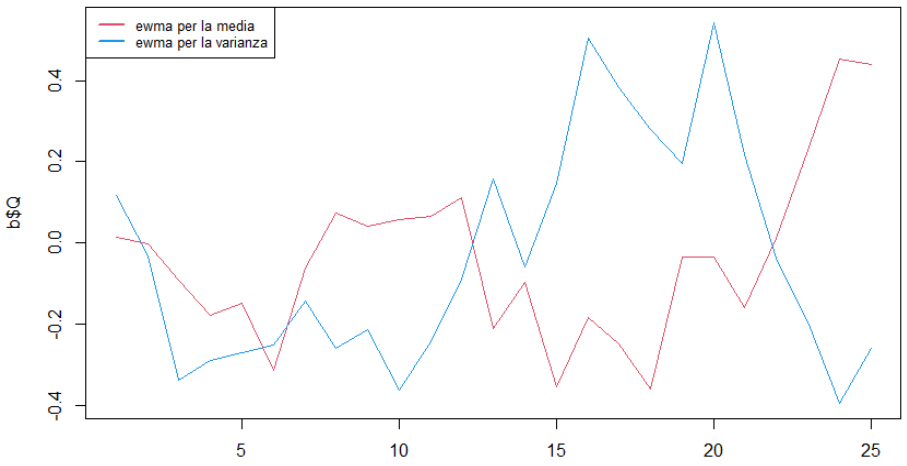
Per scrupolo, andiamo a controllare che la carta non chiami falsi allarmi, quando le forniamo i dati di fase I.

```
u2 <- matrix(Funzione_F(flow1),nrow=chart$n)
u2 = t(u2) # prepariamoci la matrice da passare a ewma.score()
b <- ewma.score(chart = chart, u = u2, plot = TRUE)
b$OC
integer(0)
```

La carta DFS non chiama allarmi per i dati di fase I, quindi possiamo considerare come affidabili i risultati ottenuti riguardo l'analisi sui dati flow2.

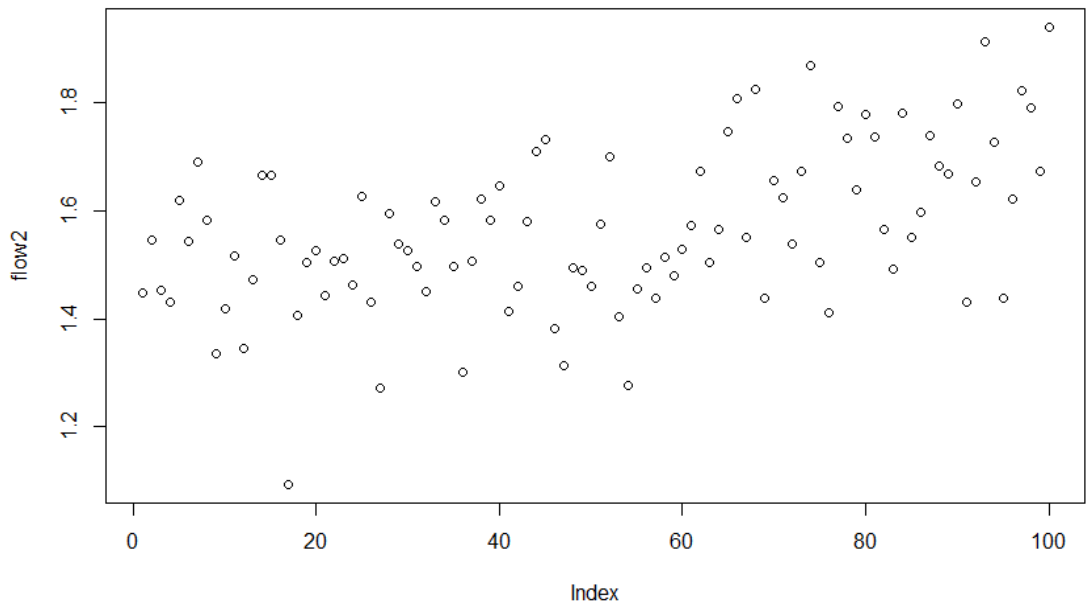


Concludo la trattazione di questo dataset riportando il grafico (poco interessante) che rappresenta l'andamento delle due componenti dello score al passare delle "infornate".



Per ulteriore conferma, si vede come la situazione di fuori controllo sia dovuta ad un aumento della media, anche semplicemente dal grafico di dispersione dei dati flow2; (si vede come dall'osservazione 60 in poi, i dati sono mediamente più grandi).

```
plot(flow2)
```



CONCLUSIONI

Questa relazione ha trattato lo studio di una carta di controllo non parametrica per sorvegliare simultaneamente la media e la varianza di un processo. La scrittura di questo elaborato si basa su un articolo degli Autori Dong Ding, Jian Li, Fugee Tsung e Yang Li (2023), nel quale viene proposta una carta di controllo non parametrica basata su una trasformazione dello score di verosimiglianza, denominata DFS (Distribution-Free chart based on the Score test).

Dopo una doverosa introduzione al mondo del controllo statistico della qualità, di cosa si occupa e quali sono gli strumenti di base che utilizza, è stata presentata una nuova carta di controllo, denominata DFS. È stato mostrato come si è arrivati a definirla e quali sono state le assunzioni necessarie per renderla effettivamente applicabile; ad esempio la scelta dell'utilizzo della distribuzione logistica come vera forma di $f(\cdot)$, al fine di avere una formulazione esplicita dello score.

Si è passati poi alla presentazione dei risultati, ottenuti via simulazione, riguardo le performance della carta DFS in termini di ARL OC, confrontandola a parità di ARL IC con altre tipologie di carte: una carta parametrica basata su una combinazione di due EWMA, una non parametrica basata sul test di bontà di adattamento (GOF) e infine una carta non parametrica basata sui ranghi. Da questi confronti si è capito che le capacità della carta DFS di chiamare un allarme rapidamente, sono molto buone e robuste (anche al variare della vera distribuzione del processo).

Nell'ultimo capitolo di questa relazione abbiamo applicato la carta a dei dati reali, per mostrare come si può utilizzare dal punto di vista pratico. Per fare ciò, abbiamo deciso, al contrario degli Autori dell'articolo, di adoperare dei limiti di controllo dinamici, i quali vengono calcolati via simulazione (una sola volta) prima di applicare la carta. Questa scelta è stata fatta sia per le garanzie che questa tipologia di limiti ci può dare in termini di probabilità di falsi allarmi, sia per la loro semplicità di calcolo.

Un secondo aspetto che gli Autori dell'articolo hanno tralasciato, ma su cui noi abbiamo voluto concentrarci, è stato quello di dare un'indicazione riguardo al motivo per cui la carta ha segnalato una situazione OC, ovvero cercare di capire se la carta sta chiamando un allarme a causa di uno scostamento (dalla situazione IC) della media, della varianza o di entrambe.

BIBLIOGRAFIA

- Ding D, Li J, Tsung F, Li Y. A Phase II score-based distribution-free method for jointly monitoring location and scale (2023), John Wiley & Sons Ltd. https://doi.org/10.1002/qre.3413
- Douglas C, Montgomery. Controllo statistico della qualità (2006), McGraw-Hill, seconda edizione.
- Peihua Qiu. Introduction to Statistical Process Control (2014), Taylor & Francis Group.
- Subhabrata Chakraborti, Marien Alet Graham. Nonparametric Statistical Process Control (2019), John Wiley & Sons Ltd.
- R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/
- Bakir ST, Reynolds MR. A nonparametric procedure for process control based on within-group ranking (1979), Technometrics.
- Hackl P, Ledolter J. A control chart based on ranks (1991), J Qual Technol.
- Chakraborti S, Eryilmaz S, Human SW. A phase II nonparametric control chart based on precedence statistics with runs-type signaling rules (2009), Comput Stat Data Analysis.
- Graham MA, Chakraborti S, Human SW. A nonparametric exponentially weighted moving average signed-rank chart for monitoring location (2011), Comput Stat Data Analysis.
- Graham MA, Mukherjee A, Chakraborti S. Distribution-free exponentially weighted moving average control charts for monitoring unknown location (2012), Comput Stat Data Analysis.
- Liu L, Tsung F, Zhang J. Adaptive nonparametric CUSUM scheme for detecting unknown shifts in location (2014), Int J Prod Res.
- Mukherjee A, Chakraborti S. A distribution-free control chart for the joint monitoring of location and scale (2012), Qual Reliab Eng Int.
- Ross GJ, Tasoulis DK, Adams NM. Nonparametric monitoring of data

streams for changes in location and scale (2011), Technometrics.

- Chowdhury S, Mukherjee A, Chakraborti S. A new distribution-free control chart for joint monitoring of unknown location and scale parameters of continuous distributions (2014), Qual Reliab Eng Int.
- Xiang D, Gao S, Li W, Pu X, Dou W. A new nonparametric monitoring of data streams for changes in location and scale via Cucconi statistic (2019), J Nonparametric Stat.
- Zou C, Tsung F. Likelihood ratio-based distribution-free EWMA control charts (2010), J Qual Technol.
- Ross GJ, Adams NM. Two nonparametric control charts for detecting arbitrary distribution changes (2010), J Qual Technol.
- Zhang J, Li E, Li Z. A Cramer-von Mises test-based distribution-free control chart for joint monitoring of location and scale (2017), Comput Ind Eng.

RINGRAZIAMENTI

In questa sezione vorrei ringraziare tutti coloro che mi hanno sostenuto nel mio percorso universitario. Ringrazio in primis i miei genitori, Roberto Passuello e Tiziana Dalla Vecchia per il notevole supporto morale ed economico.
In secondo luogo ringrazio i miei fratelli Chiara e Filippo, i numerosi parenti: nonna, zie, zii, cugini e cugine.
Ringrazio poi gli amici che ho conosciuto proprio qui al Dipartimento di Scienze Statistiche dell'Università degli Studi di Padova: persone quali Francesco Lasalvia, Filippo Scalabrin, Greta Schiappacasse, Giorgio Nagy sono e sono state fondamentali per me.
Ringrazio poi gli amici che già conoscevo, condividere dei bei momenti (quality time) insieme a loro, rende piacevoli le mie giornate. Ringrazio in particolare gli amici che apprezzano la mia grande passione per i giochi da tavolo: Alice La Barca, Alessio Palatella, Davide Lucchini, Elena Vitella, Leonardo Maniscalco, Davide Albiero, Marisol Traforetti (e soprattutto mia sorella Chiara), per nominarne qualcuno.
Altri sentiti ringraziamenti vanno anche a zia Concetta, per avermi prestato per un lungo periodo, addirittura fino ad oggi, la mitica Punto grigia del '98, con la quale mi reco alla stazione di Lerino per prendere il treno diretto a Padova.
Ringrazio mio cugino Leonardo Dalla Vecchia per avermi prestato a lungo termine il PC portatile (chiamato Pietro) con il quale ho preparato e passato alcuni esami in modalità BYOD (Bring Your Own Device).
Ringrazio tutto il personale Unipd del Dipartimento di Statistica, senza di loro non sarebbe stato così gradevole trascorrere le moltissime giornate passate a studiare in aule come la ASID60 e la ASID17.
Ringrazio i vari dottori e specialisti che, soprattutto nell'ultimo periodo, mi hanno seguito nelle due importanti operazioni agli occhi (lenti IOL per miopia e astigmatismo) e ai denti (due impianti dentali per agenesia) che ho finalmente fatto a fine 2023; in particolare il mio oculista Joseph Sajish Pinackatt, il Dottor Fabrizio Gabas di Vista Vision e il Dottor Alessio Franchina di Clinica Dentale.
Ringrazio la psicologa Laura Dalla Vecchia che mi ha seguito in un periodo critico della mia vita, il suo è stato un valido aiuto.
Ringrazio anche il personale ESU delle mense universitarie, in particolare quello della mensa Pio X, mi è sempre piaciuto pranzare da voi.
Ringrazio infine il Professore Guido Masarotto, relatore di questa Tesi e senza dubbio il miglior professore del mio percorso universitario. I giorni in cui c'era una sua lezione erano i miei preferiti, le sue spiegazioni riuscivano veramente a farmi apprezzare gli argomenti trattati, utilissimi i "casi studio" tratti dalla sua esperienza personale. Professore, concludo ringraziandola ancora, per me è stato veramente un onore averla come relatore.

BONUS

In questa sezione condivido un link e un qr-code che punta ad una cartella Google Drive contenente la Tesi in versione Google Documenti e pdf, l'articolo in cui è stata presentata la carta DFS, lo Script R utilizzato nell'elaborato e i dataset med.dat e flow.dat completi.

🗀 Tesi Federico Passuello da Condividere

https://drive.google.com/drive/folders/1S6Ac8kKEMrwMg6wRvf0l3eFpXe50sxbu?usp=sharing