PRO 5859

Statistical Process Monitoring

Linda Lee Ho

Department of Production Engineering
University of São Paulo

2020

Outline

Univariate Control chart: CUSUM & EWMA type CC

Disadvantages of Shewhart CC

- Information about the process based on the last plotted point
- Ignores any information by the entire sequence of points
- Relatively insensitive to detect small shifts as $< 1.5\ \sigma$
- Inclusion of warning limits or supplementary rules to improve the performance reduces the simplicity and easy interpretation
- Effective alternatives: CUSUM & EWMA CC

Outline

Univariate Control chart: CUSUM & EWMA type CC

- CUSUM and EWMA type CC

CUSUM CC: Introduction

- Proposed by Page (1954)
- Incorporates all information of sequential sampled values
- Plots the cumulative sum of the deviations of the sample values from a target value

$$
\begin{aligned}
C_i &= \sum_{j=1}^{i} \Theta_j - \mu_0 \\
&= (\Theta_i - \mu_0) + C_{i-1}
\end{aligned}
$$

  Θ_j is a monitored statistic
- Whenever $C_i > H$, the process is declared out-of-control
- Reasonable value for H is 5σ
- It is possible to develop CUSUM procedure for other variables than normally distributed

CUSUM CC: The statistics C_i^+, C_i^-

- Accumulating derivations above the target (from μ_0)

$$C_i^+ = max[0; C_{i-1}^+ + \Theta_i - K]$$

.

- Accumulating derivations below the target (from μ_0)

$$C_i^- = min[0; C_{i-1}^- + \Theta_i - K]$$

- K, known as **reference value** depends on the underlying distribution. For a normal distribution $K = \dfrac{|\mu_1 - \mu_0|}{2}$. Values of K for other distribution belonging to exponential family see Hawkins (1997)

CUSUM CC: Example

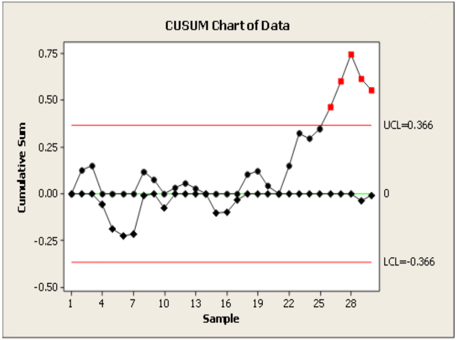


Figure 1: An example of CUSUM Chart

CUSUM CC: About ARL_0 and ARL_1

- Let $K = k\sigma$; $H = h\sigma$ and $\mu_1 = \mu_0 + \delta\sigma$
- An approximation for ARL proposed by Siegmund(1985) when X is a standard normal distribution

$$ARL \approx \frac{exp[-2(h+1.666)(\delta-k)] + 2(h+1.666)(\delta-k) - 1}{2(\delta-k)^2}$$

Thus for $\delta = 0$

$$ARL_0 \approx \frac{exp[2k(h+1.666)] - 2k(h+1.666) - 1}{2k^2}$$

And if the value of $k = 1/2$ is used then

$$ARL_0 \approx 2[exp(h+1.666) - (h+1.666) - 1]$$

CUSUM CC: About ARL_0 and ARL_1

And for a shift of $\delta = 2k$ then,

$$ARL_1 = \frac{exp[-2k(h + 1.666)] + 2k(h + 1.666) - 1}{2k^2}$$

CUSUM CC: Get h given k and ARL_0

Using 2 terms of the series expansion

$$h \approx \frac{\ln(1 + 2k^2 ARL_0)}{2k} - 1.666$$

when $k = 1/2$ h simplifies to

$$h \approx \ln\left(0.5ARL_0 + 1\right) - 1.666$$

Using 3 terms of the series expansion

$$h \approx \frac{2k^2 ARL_0 + 2}{2k^2 ARL_0 + 1} \times \frac{\ln(1 + 2k^2 ARL_0)}{2k} - 1.666$$

when $k = 1/2$ h simplifies to

$$h \approx \frac{ARL_0 + 4}{ARL_0 + 2} \times \ln(0.5ARL_0 + 1) - 1.666$$

CUSUM CC: Choice of k and h for desired ARL_0 and ARL_1

Given the desired values of ARL_0 and ARL_1 find k as

$$k = \sqrt{\frac{\ln(-\ln \nu) - \ln \nu}{2ARL_1} - \frac{1}{2ARL_0}}$$

where $\nu = \frac{ARL_1}{ARL_0} exp\left(1 - \frac{ARL_1}{ARL_0}\right)$ and then determine h using

$$h \approx \frac{2k^2 ARL_0 + 2}{2k^2 ARL_0 + 1} \times \frac{\ln(1 + 2k^2 ARL_0)}{2k} - 1.666$$

CUSUM CC: to monitor process variability

- Assumption: X_i, normally μ_0 and σ_0^2, thus $Y_i = \dfrac{X_i - \mu_0}{\sigma_0}$ is a standardized normal distribution
- Hawkins (1981,1993) suggests using $\nu_i = \dfrac{\sqrt{|y_i|} - 0.822}{0.349}$ to build a CUSUM by plotting

$$S_i^+ = max[0; \nu_i - k + S_{i-1}^+]$$

$$S_i^- = min[0; \nu_i - k + S_{i-1}^-]$$

as the in-control distribution of ν_i is approximately $N(0; 1)$

- Values of k and h are selected to meet some performance metric

CUSUM CC: Questions for Seminars

- It is possible to develop CUSUM for other statistics.
- In Olwell & Hawkins(1998), they affirmed: *"This general method of finding CUSUM schemes for changes in a parameter may be used for any member of the exponential family of distributions. The CUSUM is of the sufficient statistic minus a reference value*".
- Discuss which are CUSUM schemes for some distributions from the exponential family distribution
- How to find the values of k and h in these cases?
- Suggestion: Olwell & Hawkins, (1998) Cumulative sum charts and charting for quality improvement - Chapter 6
- Good reference about CUSUM:Hawkins (1997)

EWMA type control chart: Introduction

- Proposed by Roberts (1959)
- Definition

$$Z_i = \lambda X_i + (1-\lambda) Z_{i-1}$$

$0 \leq \lambda \leq 1$, a constant with $Z_0 = \mu_0$

- Writing recursively

$$Z_i = \lambda \sum_{j=0}^{i-1} (1-\lambda)^j X_{i-j} + (1-\lambda)^i Z_0$$

The weights decrease geometrically with

$$\lambda \sum_{j=0}^{i-1} (1-\lambda)^j = 1 - (1-\lambda)^i$$

EWMA-type control chart: Introduction

- If the observations X_j are independent with variance σ^2, then

$$Var(Z_i) = \sigma^2 \left(\frac{\lambda}{2-\lambda} \right) [1 - (1-\lambda)^{2i}]$$

- Control Limits:

$$\mu_0 \pm L\sigma \sqrt{\frac{\lambda}{2-\lambda}[1 - (1-\lambda)^{2i}]}$$

L, searched to reach a desired performance

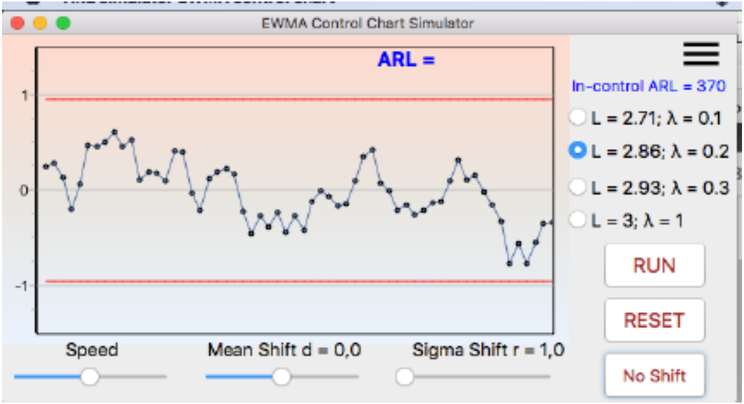
- $[1 - (1-\lambda)^{2i}] \rightarrow 1$ as $i \rightarrow \infty$ then steady-state control limits:

$$\mu_o \pm L\sigma \sqrt{\frac{\lambda}{2-\lambda}}$$

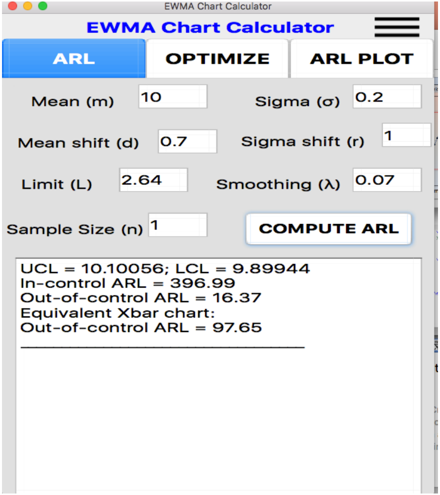
EWMA-type control chart: performance, λ and L

- EWMA properly designed may be used to monitor individual observations even for non-normality distributions
- EWMA performs well against small shifts
- Often superior for CUSUM for large shifts
- Combination of EWMA and Shewhart charts effective against both large and small shifts
- Select λ and L to meet desired ARL performance. Some rules of thumbs:
  - $\lambda \in [0.05; 0.25]$ work well in practice
  - Smaller values of λ to detect small shifts
  - $L = 3$ works reasonably well for large values of λ

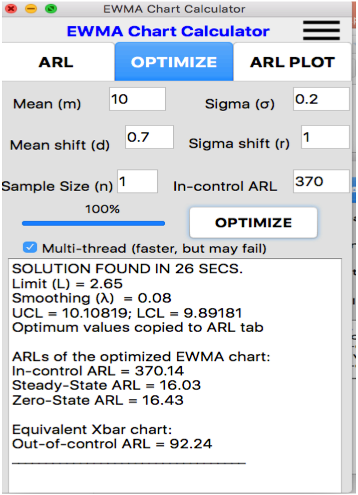
EWMA-type control: APP available



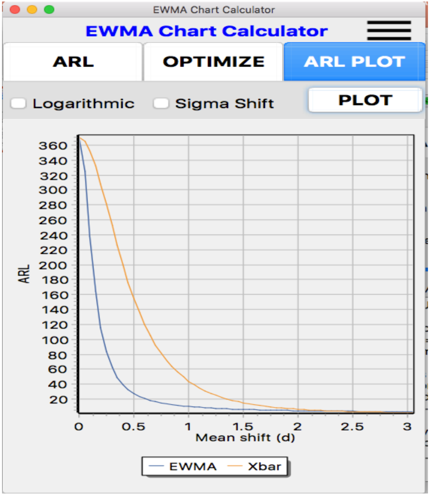
EWMA-type control: APP available



EWMA-type control: APP available



EWMA-type control: APP available



EWMA-type control chart: to monitor variability

- X_i normally distributed with μ and σ^2. Let

$$S_i^2 = \lambda(X_i - \mu)^2 + (1-\lambda)S_{i-1}^2$$

For X_i independent observations, S_i^2 follows approximately chi-square distribution with d.f $\nu = \dfrac{2-\lambda}{\lambda}$.
As $E(S_i^2) = \sigma^2$ for large i thus S_i^2 or S_i can be plotted to build the EWMA chart.
If S_i is plotted, the control limits are:

$$UCL = \sigma_0\sqrt{\frac{\chi^2_{\nu;\alpha/2}}{\nu}}; LCL = \sigma_0\sqrt{\frac{\chi^2_{\nu;1-\alpha/2}}{\nu}}$$

EWMA-type control chart: to monitor variability

- MacGregor and Harris(1993) suggests replacing μ by an estimate $\hat{\mu}_i$ at each time
- Logical estimate of μ is z_i, thus

$$S_i^2 = \lambda(X_i - z_i)^2 + (1-\lambda)S_{i-1}^2$$

can be used to build EWMA CC

EWMA-type control chart: for Poisson data

if X_i is a count then

$$z_i = \lambda x_i + (1-\lambda) z_{i-1}$$

can be used to build a EWMA CC having control limits as

$$UCL = \mu_0 + A_U \sqrt{\frac{\lambda \mu_0}{2-\lambda}[1-(1-\lambda)^{2i}]}$$

$$LCL = \mu_0 + A_L \sqrt{\frac{\lambda \mu_0}{2-\lambda}[1-(1-\lambda)^{2i}]}$$

A_i, i=U,L are constants to be searched. Many applications use $A_U = A_L = A$

EWMA CC: Questions for Seminar

- Find in the literature other applications/ways/statistics for EWMA

Application of CUSUM chart

- CUSUM Control charts to monitor series of Negative Binomial count data (Alencar et al. 2017)
- Illustration: The monitoring of daily number of hospital admissions due to respiratory diseases for people aged over 65 years old in the city São Paulo-Brazil
- Two step procedure:
  1. A generalized linear model (GLM) is fitted for data - Jan 2006-Dec 2010;
  2. Control chart applied for data of 2011

Assumptions

- Only a single observation X_t is considered at each time t and X_t assumes non negative integer values in $\{0, 1, \ldots, \}$;
- The expected value may vary over time as a function of explanatory variables;
- X_t follows a Poisson or Negative Binomial distribution.

Daily admissions - respiratory diseases

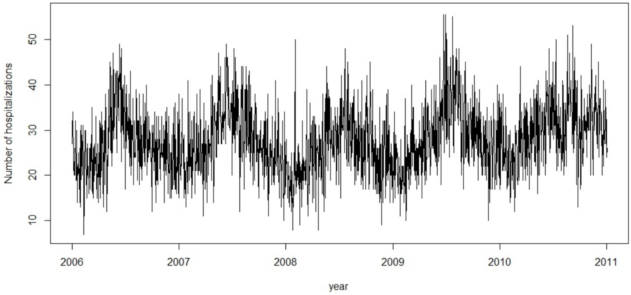


Figure 6: Daily admissions due to respiratory diseases for people aged over 65 in São Paulo from 2006 to 2011

Features in GLM

- The daily hospitalization series presents a seasonal behavior;
- To control part of this seasonality, sine and cosine functions are included in GLM;
- Days of week included as categorical variables to explain the variability of daily hospitalizations.

Fitting GLM

- X_t = the daily number of hospitalizations at day t following $\sim NegBin(\mu_{0,t}, \phi)$
- Its expectation $\mu_{0,t}$ written as

$$
\begin{aligned}
ln\left(\frac{\mu_{0,t}}{pop_t}100000\right) &= \beta_0 + \beta_1 cos(2\pi t/365) + \beta_2 sin(2\pi t/365) \\
&+\beta_3 Sat_t + \beta_4 Sun_t + \beta_5 Mon_t,
\end{aligned}
$$

- pop_t, the population in risk on the t-th day;
- Sat_t, Sun_t and Mon_t, dummy variables equal to 1 respectively for Saturday, Sunday, and Monday, and zero otherwise.
- The inclusion of the offset term $g_t = (pop_t/100000)$ allows to model the average daily rate of admissions

Fitting GLM

Table 1: Estimates, standard errors and p-values of the parameters of model (1)

Coefficient	Estimate	S.E.	p-value
Intercept (β_0)	1.227	0.007	< 0.001
Cossine (β_1)	-0.173	0.007	< 0.001
Sine (β_2)	-0.045	0.007	< 0.001
Saturday (β_3)	-0.158	0.016	< 0.001
Sunday (β_4)	-0.207	0.016	< 0.001
Monday (β_5)	0.040	0.015	0.008
Dispersion (ϕ)	69.99	8.22	

Fitting GLM

- The fitness of the proposed model evaluated by a residual analysis.
- The residual deviance is 1847.7 for 1825 degrees of freedom, which is an indicative of good fit.
- The usual residuals plots indicated that there is no departure of the assumptions of the model, excepted the independence of the residuals, since the autocorrelations of the deviance residuals are around 0.2 for the first and second lags.

Fitting GLM

- QQ-plot of the deviance residuals in Figure 7; it seems they are normally distributed;

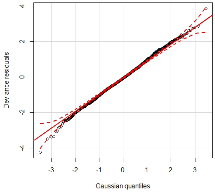


Figure 7: Quantile-quantile plot of deviance residuals

- The normality assumption rejected by the Shapiro Wilk test (p=0.1972), thus the assumption of a Negative Binomial distribution for the counts is appropriate.
- No evidence of outliers among the observations.

Observed versus Predicted

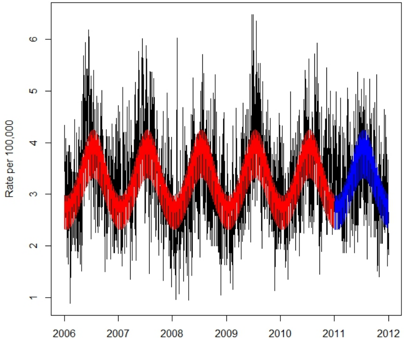


Figure 8: Observed daily admissions in São Paulo from 2006 to 2011 and predicted values

Monitored Statistics (# 1)

$$Z^*_{0,t} = \frac{X_t - n_t\mu_t}{\sqrt{n_t\mu_t}}, \quad t = 1, 2, \ldots,$$

$$Z^{**}_{0,t} = 2(\sqrt{X}_t - \sqrt{n_t\mu_t}), \quad t = 1, 2, \ldots,$$

- $$Z_{1,t} = 0.5Z^*_{0,t} + 0.5Z^{**}_{0,t} = \frac{X_t - 3n_t\mu_t + 2\sqrt{X_t n_t\mu_t}}{2\sqrt{n_t\mu_t}}.$$

where $n_1, n_2, \ldots$ are the corresponding sample sizes with $E(X_t) = n_t\mu_t$.

Monitored statistics (#2, #3)

- $$Z_{2,t} = \sqrt{\phi - a}\left(\sinh^{-1}\sqrt{\frac{X_t + b}{\phi - 2b}} - \sinh^{-1}\sqrt{\frac{\mu_t + b}{\phi - 2b}}\right),$$

  with $a = b = 0$ or $a = 0.5; b = 0.375$.
- $$Z_{3,t} = \sqrt{\phi - 0.5}\left(\sqrt{\frac{X_t + 0.385}{\phi - 0.75}} - \sqrt{\frac{\mu_t + 0.385}{\phi - 0.75}}\right).$$

Monitored Statistics (#4, #5)

- $$Z_{4,t} = \frac{X_t - \mu_t}{\sqrt{\phi\pi_t/(1-\pi_t)^2}},$$
  with $\pi_t = \mu_t/(\mu_t + \phi)$.
- $$Z_{5,t} = sign(X_t - \mu_t)\sqrt{(d_t^2)},$$
  where
  $$d_t^2 = \begin{cases} 2\phi \ln(1 + \mu/\phi), & if\ X_t = 0 \\ 2X_t \ln\left(\frac{X_t}{\mu_t}\right) - 2\phi(1 + X_t/\phi)\ln\left(\frac{1+X_t/\phi}{1+\mu_t/\phi}\right) & if\ X_t > 0 \end{cases}$$
  ϕ is a constant dispersion parameter, $Z_{5,t}$ follows approximately a standardized normal distribution.

CUSUM charts #1 to #5

- $$C_{i,t} = max(0, C_{i,t-1} + Z_{i,t} - k_t), i = 1, \ldots, 5.$$

  Whenever $C_{i,t} > h$, it is decided that the process is out-of-control, meaning that the monitored parameter has shifted and a search for special causes starts.

CUSUM charts #6 , #7

- $$C_{6,t} = max[(0, C_{6,t-1} + c_t(X_t - k_t))]$$

  with k_t determined under a Negative Binomial distribution as

  $$k_t = \frac{-\phi \ln \{(\phi + \mu_{0,t})/(\phi + \mu_{1,t})\}}{\ln \{\mu_{1,t}(\phi + \mu_{0,t})/\mu_{0,t}(\phi + \mu_{1,t})\}},$$

- $$C_{7,0} = 0, \quad C_{7,t} = max\left(0, C_{7,t-1} + ln\left\{\frac{f_{\theta_0}(x_t)}{f_{\theta_1}(x_t)} - k\right\}\right), \quad t \geq 1$$

Performance of CUSUM control chart

- The reference value k and the control limit h of CUSUM chart are searched by simulation;
- Count time series $X_{i,t}$ with $\mu_{0,t}$ are simulated
- The cumulative sums $C_{i,t}$ are calculated
- 10,000 simulated run lengths are used to meet $ARL_0 = 500$

Performance of CUSUM control chart

Table 2: Average Run Length and standard errors (SE) for CUSUM chart - JK

			$\delta = 1.0$		$\delta = 1.25$		$\delta = 1.5$		$\delta = 1.75$		$\delta = 2.0$	
k	h	ARL_0	SE	MLR_0	ARL_1	SE	ARL_1	SE	ARL_1	SE	ARL_1	SE
0.0	13.47	500.71	4.13	373.00	20.40	0.05	10.65	0.02	7.32	0.01	5.74	0.01
0.1	7.14	500.77	4.36	359.00	13.01	0.04	6.39	0.02	4.43	0.01	3.54	0.01
0.2	4.94	499.79	4.35	355.50	10.94	0.05	5.06	0.01	3.50	0.01	2.80	0.01
0.3	3.80	500.21	4.35	362.00	10.87	0.05	4.60	0.01	3.13	0.01	2.48	0.01
0.4	3.07	500.25	4.28	371.50	**10.66**	0.06	4.11	0.01	2.74	0.01	2.15	0.01
0.5	2.57	499.85	4.29	365.00	11.56	0.08	3.94	0.02	2.55	0.01	1.95	0.01
0.6	2.20	500.00	4.28	366.50	12.92	0.10	**3.87**	0.02	2.44	0.01	1.84	0.01
0.7	1.92	499.73	4.26	376.00	14.82	0.12	3.90	0.02	2.35	0.01	1.74	0.01
0.8	1.69	500.81	4.31	359.00	17.36	0.15	4.00	0.02	2.29	0.01	1.65	0.01
0.9	1.51	500.85	4.33	364.00	20.17	0.17	4.16	0.02	2.31	0.01	1.65	0.01
1.0	1.35	500.17	4.37	352.00	23.11	0.20	4.38	0.03	**2.29**	0.01	**1.61**	0.01
1.1	1.21	499.36	4.35	357.00	26.33	0.23	4.67	0.03	2.33	0.01	1.62	0.01

Performance of CUSUM chart

Table 3: The "best" design parameters (k and h) which provides the best performance in terms of ARL_1

		$\delta = 1.25$			$\delta = 1.5$			$\delta = 1.75$			$\delta = 2.0$		
		k	h	ARL_1	k	h	ARL_1	k	h	ARL_1	k	h	ARL_1
$C_{1,t}$	RS	0.6	5.28	9.30	1.1	3.20	3.47	1.5	2.37	2.02	1.6	2.22	1.46
$C_{2,t}$	JK	0.4	3.07	10.66	0.6	2.20	3.87	1.0	1.35	2.29	1.0	1.35	1.61
$C_{3,t}$	GN	0.3	2.42	9.43	0.5	1.56	3.50	0.8	0.96	2.05	1.0	0.70	1.43
$C_{4,t}$	JG	0.5	4.74	8.76	0.9	2.97	3.25	1.2	2.30	1.89	1.2	2.30	1.39
$C_{5,t}$	DR	0.5	4.10	8.84	0.9	2.46	3.27	1.2	1.84	1.90	1.2	1.84	1.39
$C_{6,t}$	RY	-8.5	26.4	**7.18**	-5.1	15.64	**3.07**	-3.7	13.3	**1.88**	-3.7	13.3	**1.39**
$C_{7,t}$	LR	-3.8	11.90	**7.35**	-3.1	8.92	**3.02**	-1.5	5.56	**1.80**	-1.5	5.56	**1.34**

CUSUM Chart for 2011

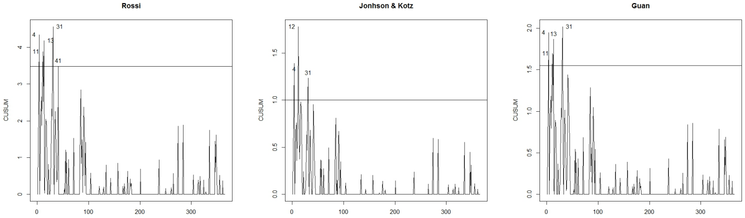
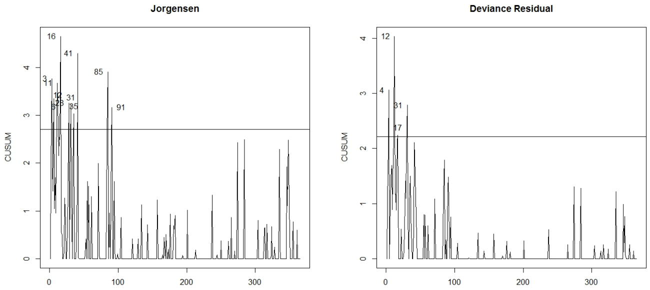


Figure 9: CUSUM charts for Hospitalizations in 2011 for $ARL_0 = 500$ and $\delta = 1.5$

CUSUM Chart for 2011



CUSUM Chart for 2011

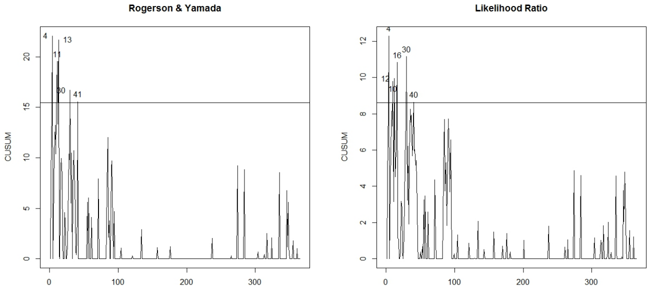


Figure 10: CUSUM charts for Hospitalizations in 2011 for $ARL_0 = 500$ and $\delta = 1.5$

Alencar, A. P., Lee Ho, L. & Albarracin, O. Y. E. (2017), 'Cusum control charts to monitor series of negative binomial count data', *Statistical methods in medical research* **26**(4), 1925–1935.

Hawkins, D.M., O. D. (1997), *Cumulative Sum Charts and Charting for Quality Improvement*, Springer.