CUSUM Burst Detection in High Frequency Markets

W.L. Gore Statistician/Data Scientist Interview

Ivan E. Perez

November 2, 2022

Presentation Outline

Background and Motivation

As a leading manufacturer of vents and ePTFE products, W.L. Gore and Associates (Gore) employs **s**tatistical **p**rocess **c**ontrol (SPC) methods, to improve quality control and detect when production processes may begin to impact vent quality. [1, 2]

This work:

1. Outlines a dual **Cu**mulative **SUM**, 2-**CUSUM** detection scheme for a time-**H**omogeneous **C**ompound **P**oisson **P**rocess (HCPP).
2. Shows how 2-CUSUMs can be used detect when a HCPP is out of control.
3. Demonstrates how a sequence of 2-CUSUM alarms can be used to identify out of control burst periods.
4. Analyze the effect these burst periods have on momentum, (i.e., price range/unit time)

This work can be beneficial to Gore in detecting changes in the defect rate of a continuous production processes where defects

- arrive randomly and in short bursts, and
- can be quantified using continuous or categorical **r**andom **v**ariables, RV.

The BTC-USD Coinbase Exchange

The Coinbase Exchange is a transparent exchange that publicly lists all market, limit, orders and the instantaneous state of the limit Order Book.[3]

This work focuses on **Market Orders**; a buy or sell order that is immediately filled at best price and size of the market making side.

Table: Six recorded market order transactions collected from the Coinbase BTC-USD ticker channel using CoPrA[4]

Timestamp (UTC)	Volume (BTC)	side	Price (USD)
19/03/2020 22:10:46	0.015439	buy	6252.08
19/03/2020 22:10:47	0.494871	buy	6255.91
19/03/2020 22:10:49	0.017539	buy	6254.72
19/03/2020 22:10:51	0.007641	buy	6254.73
19/03/2020 22:10:52	1.061969	sell	6254.72
19/03/2020 22:10:55	0.069624	buy	6255.90

Modeling Market Order Arrivals as a time-Homogeneous Compound Poisson Process (HCPP)

Definition: time-Homogeneous Compound Poisson Point Process[5]

A stochastic process $\{Q(t)\}_{t\geq 0}$ is said to a compound Poisson process if it can be represented as:

$$Q(t) = \sum_{i=1}^{N(t)} Y_i, \qquad t \geq 0$$

where $\{N(t)\}_{t\geq 0}$ is a Poisson process, and $\{Y_i\}_{i\geq 1}$ is a family of i.i.d. RV's independent of $\{N(t)\}_{t\geq 0}$

We can apply Wald's equality to $Q(t)$:

$$\mathbb{E}[Q(t)] = \mathbb{E}\left[\sum_{i=1}^{N(t)} Y_i\right] = \mathbb{E}[N(t)]\mathbb{E}[Y_1] = \lambda t \mu, \tag{1}$$

In other words in a interval length t, the expected sum of arrivals $\lambda \mu t$.

Modeling Market Order Arrivals as a time-Homogeneous Compound Poisson Process (HCPP)

1. Modeling Market Order arrivals as a HCPP with message rate λ(msg/s), and mean order size, μ(BTC/msg).
2. Therefore, in an interval of length t, with $N(t)$ number of messages, and orders with mean μ, the expected total transaction volume in the interval is $\lambda\mu t$.
3. In order to estimate the baseline order arrival:
   - *rate* $\hat{\lambda}$, we count X messages in an interval of length t,
   - *size* $\hat{\mu}$, we take the mean size of messages in the interval of length t.

Applying our HCPP model our orders can be described as:

Timestamp (UTC)	Volume (BTC) $= Y_i$	t (s)	$N(t)$	$Q(t)$
19/03/2020 22:10:45	0.000000	0	0	0
19/03/2020 22:10:46	0.015439	1	1	0.015439
19/03/2020 22:10:47	0.494871	2	2	0.649261
19/03/2020 22:10:49	0.017539	4	3	0.666800
19/03/2020 22:10:51	0.007641	6	4	0.674441
19/03/2020 22:10:52	1.061969	7	5	1.736410
19/03/2020 22:10:55	0.069624	10	6	1.806034

MLE of Expected Size of Market Order Arrivals in a period length t, $\lambda t \mu$

For a sequence of X messages in n periods of length t, $X_1, \ldots X_n$, our **M**aximum **L**ikelihood **E**stimator (MLE) for the rate of order arrival is

$$\hat{\lambda} = \frac{1}{nt} \sum_{i=1}^{n} X_i = \frac{N(nt)}{nt}, \qquad n = 1, 2, \ldots \tag{2}$$

For the size of the market orders, recognize that their size is time invariant, but is indexed by $N(t)$. We take their mean at time nt to be:

$$\hat{\mu} = \frac{1}{N(nt)} \sum_{i=1}^{N(nt)} Y_i. \tag{3}$$

To assess whether the data obtained in 60 non-overlapping minute volume intervals follows a HCPP model in line with MLE rate $\hat{\lambda}\hat{\mu}$ we perform a χ^2-test with hypotheses:

$$H_0 : X_1, \ldots, X_{N(t)} \sim \text{HCPP with mean } \hat{\lambda}\hat{\mu}t.$$
$$H_1 : X_1, \ldots, X_{N(t)} \nsim \text{HCPP with mean } \hat{\lambda}\hat{\mu}t.$$

χ^2-test Results for HCPP of Market Order Arrival Rate

Table: First and last entries of the χ^2 table to test the volume of market orders in 60 non-overlapping 1-minute intervals against an expected constant rate left, $\hat{\lambda}\hat{\mu} = 5.8483$ (BTC/min), right, $\hat{\lambda}\hat{\mu} = 2.7089$ (BTC/min).

	O	E	$\frac{(O-E)^2}{E}$
X_1	0.0346	5.8483	5.7793
X_2	2.4703	5.8483	1.9511
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{59}	1.3989	5.8483	3.3852
X_{60}	10.1915	5.8483	3.2253

	O	E	$\frac{(O-E)^2}{E}$
X_1	4.3675	2.7089	1.0155
X_2	5.1747	2.7089	2.2445
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{29}	4.2100	2.7089	0.8317
X_{30}	1.6771	2.7089	0.3931

Table: χ^2-test Results

Table	df	χ^2_{crit}	χ^2_{calc}	Reject H_0 ?
Left	59	77.931	352.01	Yes
Right	29	42.56	41.50	No

The Sequential Probability Ratio Test (SPRT)

Suppose at some unknown time, $\nu, \nu = 1, 2, \ldots \leq N(t)$, when our sequence of market order arrivals, $Y_1, Y_2, Y_3, \ldots, Y_{N(t)}$ goes from a base rate $\lambda_0\mu$ to an increased rate of arrival $\lambda_1\mu$.

To identify a stopping time τ, for the detection of the change $\lambda_0\mu \to \lambda_1\mu$, we use the log Sequential Probability Ratio Test, log SPRT, $\ell(Y_1, \ldots, Y_{N(t)})$, with boundaries $(0, h)$. Our simple hypotheses are:

$$H_0 : Y_1, \ldots, Y_{N(t)} \sim \text{HCPP w/ rate } \lambda_0\mu,$$
$$H_1 : Y_1, \ldots, Y_{N(t)} \not\sim \text{HCPP w/ rate } \lambda_0\mu.$$

Our test becomes

$$\ell(Y_1, \ldots, Y_{N(t)}) \geq h \Rightarrow \text{Reject } H_0,$$
$$\ell(Y_1, \ldots, Y_{N(t)}) \leq 0 \Rightarrow \text{Fail to Reject } H_0,$$
$$\ell(Y_1, \ldots, Y_{N(t)}) \in (0, h) \Rightarrow \text{Continue sampling.}$$

Taking $\ell(Y_1, \ldots, Y_{N(t)}) = u_t(\lambda_0, \mu)$ our stopping time

$$\tau = \inf\{t \geq 0; u_t(\lambda_0, \mu) \notin (0, h)\} \tag{4}$$

i.e., the *earliest* time t where the log SPRT statistic, u_t is not between 0 and h.

CUSUM up and down definitions & application for our work

Definition: 2-CUSUM Construction

Let $\{Q(t)\}_{t\geq 0}$ be a compound Poisson process with mean $\lambda_0 \mu t$, where $\lambda_1^+ = (1+\epsilon)\lambda_0$, $\lambda_1^- = (1-\epsilon)\lambda_0$, $\{h, \lambda_0, \epsilon, \mu\} \in \mathbb{R}^+$ define the following processes:

$$u_t^+(\lambda_0,\mu) = Q(t)\log\left[\frac{\lambda_1^+}{\lambda_0}\right] - \lambda_0\epsilon t; m_t^+(\lambda_0,\mu) = \inf_{0\leq s\leq t} u_s^+(\lambda_0,\mu) \tag{5}$$

$$u_t^-(\lambda_0,\mu) = Q(t)\log\left[\frac{\lambda_1^-}{\lambda_0}\right] + \lambda_0\epsilon t; m_t^-(\lambda_0,\mu) = \inf_{0\leq s\leq t} u_s^-(\lambda_0,\mu) \tag{6}$$

$$y_t^+(\lambda_0,\mu) = u_t^+(\lambda_0,\mu) - m_t^+(\lambda_0,\mu), \tag{7}$$

$$y_t^-(\lambda_0,\mu) = u_t^-(\lambda_0,\mu) - m_t^-(\lambda_0,\mu), \tag{8}$$

$$\tau = \inf\{t\geq 0; y_t^+(\lambda_0,\mu) \vee y_t^-(\lambda_0,\mu) \geq h\}, \leftarrow \text{the CUSUM stopping time.} \tag{9}$$

Classification of Stopping times (alarms), τ

To identify shift to an increased(decreased) rate, $\lambda_0 \to \lambda_1^+$; $(\lambda_0 \to \lambda_1^-)$ we distinguish τ:

$$\tau^+ = \inf\{t\geq 0; y_t^+(\lambda_0,\mu) \geq h\}, \text{an upward alarm is declared.}$$

$$\tau^- = \inf\{t\geq 0; y_t^-(\lambda_0,\mu) \geq h\}, \text{a downward alarm is declared.}$$

Implementation of 2-CUSUM Construction - Short Experiments

Applying 2-CUSUM construction to our sequence of market orders, with an initial estimate for λ_0 we display CUSUM-up and CUSUM-down.

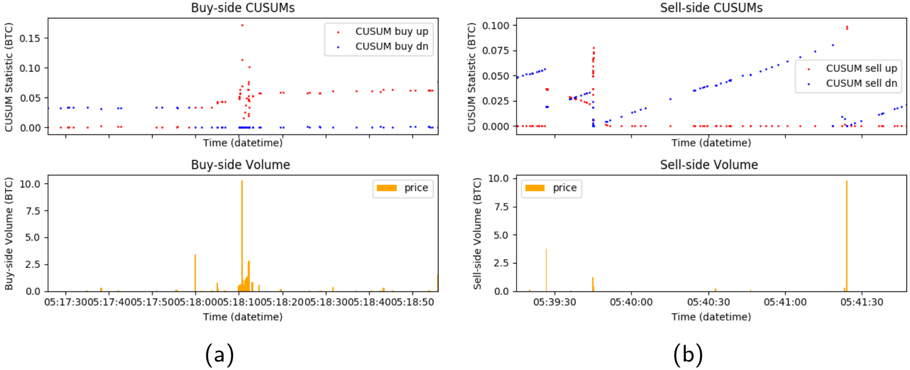


Figure: Two-sided CUSUM processes for buy(a) and sell(b) sides, under set with last transaction volume (BTC)

Results show *short bursts of high transaction volume* that trigger many upward alarms in *quick succession*.

Using a Sequence of CUSUM alarms to define Active/Inactive Periods

Goal: Classify Market Orders into Active/Inactive Periods

For a set $\mathcal{O}$ of $i, i = 1, 2, \ldots N(t)$ market orders, with a Price P_i, size, Y_i, and time t_i, our aim is to classify them into sets of Active periods, $\mathcal{A}$, and Inactive periods, where $\mathcal{I} = \{\mathcal{A}^c \in \mathcal{O}\}$.

Definition: Active/Inactive Periods

An **Active** period, $A_n, n = 1, 2, \ldots \in \mathcal{A}$ contains the set of markets orders, $\{(S_1, Y_1, t_1), \ldots, (S_i, Y_i, t_i)\}$, that contain at least 2 or more upward alarms. **Inactive** periods, $I_n, n = 1, 2, \ldots \in \mathcal{I}$, are the remaining orders.

Active Period Start and End Points

An Active periods begins on τ_1^+, and continues if there is a second up-alarm, τ_2^+ shortly thereafter. Let $T = \tau_j^+ - \tau_{j-1}^+, j = 1, 2, \ldots$.

An Active period ends on the detection of a down-alarm, τ^- or if T seconds have passed since the last up-alarm, τ_j.

Active Period Start Compensation

In order to roughly capture ν, we shift the beginning of the active period backwards by $\min[D, T]$. Where $D = \tau_1^+ - t_0$, represents time between last reset and the first alarm. t_0 is the time y_t^+ was last reinitialized.

Defining Active/Inactive Periods - Short Experiments

Applying our definition for Active(red) and Inactive(green) periods to results from the 2-CUSUM construction.

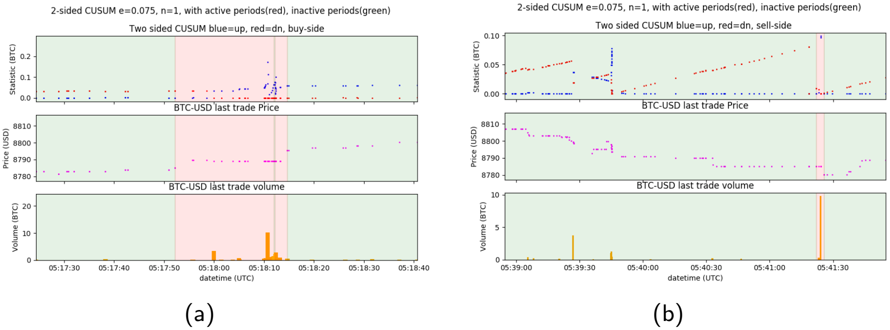


Figure: Two-sided CUSUM processes for buy(a) and sell(b) sides, under set with last transaction volume (BTC)

Results show that sometimes there are larger swings in BTC-USD Price(magenta) during Active periods.

Defining Price-Swings as Momentum

Defintion: Momentum, m

For an Active/Inactive Period P, of length t, we take momentum

$$M = \frac{\max_{i \in P} S_i - \min_{i \in P} S_i}{t} \tag{10}$$

as the range of price, divided by the length of the period in t seconds.

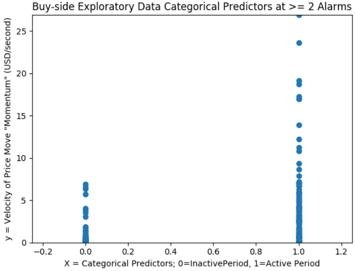
Linear Model

By encoding Active periods as 1, and Inactive periods $= 0$. The set of m momentum observations becomes $\{(M_i, E_i), i = 1, 2, \ldots m\}$. The linear model

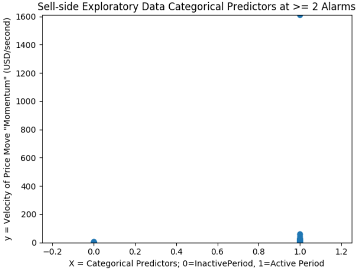
$$M = \beta_0 + \beta_1 E$$

examines the level effect β_1 on M from categorical predictor E.

24h Experiments - 2 alarms



(a)

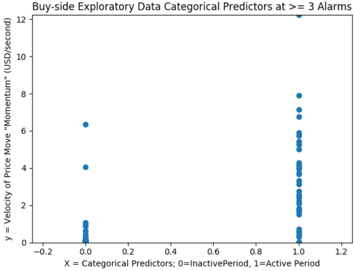


(b)

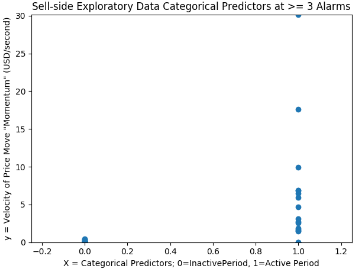
Figure: Categorical data for active(1), and inactive(0) periods against momentum for buy(a), and sell(b) side momentum experiments with at least 2 consecutive up alarms

Summary 2-alarms, buy-side				
Groups	Count	Sum	Average	VAR.S
0	124	200.986	0.5422	1.621
1	123	2944.891	3.986	23.942
Summary 2-alarms, sell-side				
Groups	Count	Sum	Average	VAR.S
0	91	151.812	0.479	1.668
1	90	2.551 E 06	25.682	2.834 E 04

24h Experiments - 3 alarms



(a)



(b)

Figure: Categorical data for active(1), and inactive(0) periods against momentum for buy(a), and sell(b) side momentum experiments with at least 2 consecutive up alarms

Summary 3-alarms, buy-side				
Groups	Count	Sum	Average	VAR.S
0	41	53.204	0.417	1.299
1	40	281.868	2.933	7.047
Summary 3-alarms, sell-side				
Groups	Count	Sum	Average	VAR.S
0	20	0.350	0.115	0.018
1	19	0.989	5.550	51.631

One-Way ANOVA of Active & Inactive Periods

The One-Way ANOVA tests the Hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. for the two groups of observed Momentums.

Table: ANOVA Table for Buy/Sell side experiments with at least 2 and 3 alarms.

cat.	Source of Variation	SS	df	MSS	F	F-crit
buy 2-alarms	Between Groups Within Groups Total	732.445 3145.878 3878.322	1 245	732 12.840	57.043	3.877
	Source of Variation	SS	df	MSS	F	F-crit
sell 2-alarms	Between Groups Within Groups Total	2.874 E 04 2.551 E 06 2.580 E 06	1 179	0.479 25.682	2.017	3.894
	Source of Variation	SS	df	MSS	F	F-crit
buy 3-alarms	Between Groups Within Groups Total	128.216 335.109 463.325	1 79	2.874 E 04 1.425 E 04	30.226	3.962
	Source of Variation	SS	df	MSS	F	F-crit
sell 3-alarms	Between Groups Within Groups Total	277.266 981.339 1258.604	1 37	128.216 4.242	10.454	4.105

Conclusions

Conclusions:

1. Rates of market order arrivals can be estimated locally using a HCPP model and its associated MLE's.
2. 2-CUSUM detection schemes can be used to detect changes to increased(decreased) rates of market order arrival.
3. Sequences of alarms can be used to identify short periods of with large price swings.
4. These short period have are known statistically significant effect sizes.

Applications for W.L. Gore

Suppose gore sells ePTFE tape, with a guarantee that 97% of the film's length falls within 0.5mm $\pm$ 0.030 mm.
To guarantee this we use a detector[6] that measures film thickness for a continuous length of ePTFE tape in intervals of length t.
The extruder/roller moves at a constant speed so the length of tape covered in a period t seconds long, is s cm, (e.g., 1 meter of tape can be extruded and rolled in 5 seconds).
The detector/recorder will:

1. Instantaneously record when the film thickness is above/below the 0.030mm threshold, and then
2. note of the time and size of the deviation above the threshold.

Under this construction the random arrival and random size of defects allows the possibility of begin modeled as a HCPP!
We can then:

1. classify periods where the tape is out of control, and
2. reject once the ratio of tape is $<$97% in control.

Appendix: Maximum Likelihood Estimation (MLE) of λ[7]

Recall the probability mass function (PMF) of a discrete Poisson random variable X with mean λt:

$$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \qquad k = 0, 1, 2, \dots \tag{11}$$

The joint distribution of $X_1, X_2, X_3, \dots X_n$ is the product of their PMFs.

$$L(X_1, X_2, \dots, X_n; \lambda) = \prod_{i=1}^{n} \frac{(\lambda t)^{X_i} e^{-\lambda t}}{X_i!}. \tag{12}$$

Taking the log of the likelihood function,

$$l(X_1, X_2, \dots, X_n; \lambda) = \sum_{i=1}^{n} \left(X_i \log \lambda t - \lambda t - \log X_i! \right). \tag{13}$$

Differentiate with respect to λ and set equal to 0 to afford

$$\frac{1}{\lambda} \sum_{i=1}^{n} X_i - nt = 0. \tag{14}$$

Solving for λ and recognizing that the running count of messages is $N(t)$, our MLE estimator is

$$\hat{\lambda} = \frac{1}{nt} \sum_{i=1}^{n} X_i = \frac{N(nt)}{nt}, \qquad n = 1, 2, \dots \tag{15}$$

Appendix: Highlights of the Probability Ratio Test[8]

Likelihood Ratio Test & Neymann Pearson Lemma

When testing a random variable X, to determine whether it is distributed under the PMF, $f(X;\theta_0)$, or $f(X;\theta_1)$, we recall the Neymann-Pearson lemma, and test whether their likelihood ratio, $L(x) = f(X;\theta_1)/f(X;\theta_0)$ exceed some constant r. Our test becomes:

$$L(x) \geq r \Rightarrow \text{Reject } H_0,$$
$$L(x) \leq r \Rightarrow \text{Fail to Reject } H_0,$$

Sequential Probability Ratio Test(SPRT)

To test whether a potentially infinite sequence of random variables $X_1, \ldots, X_n$ follows $f(x;\theta_0)$ or $f(x;\theta_1)$, we perform a sequence of probability ratio tests where ratio is $\mathcal{L}(X_1,\ldots,X_n) = f(X_1,\ldots,X_n;\theta_1)/f(X_1,\ldots,X_n;\theta_0)$, choosing positive constants $A < B$ our test becomes:

$$\mathcal{L}(X_1,\ldots,X_n) \geq B \Rightarrow \text{Reject } H_0,$$
$$\mathcal{L}(X_1,\ldots,X_n) \leq A \Rightarrow \text{Fail to Reject } H_0,$$
$$\mathcal{L}(X_1,\ldots,X_n) \leq (A,B) \Rightarrow \text{Continue sampling.}$$

We **stop** sampling when first random time N, such that $\mathcal{L}(X_1,\ldots,X_n) \notin (A,B)$.

Appendix: Derivation of the log SPRT process, $u_t(\lambda_0, \mu)$[9]

Starting from the SPRT, $\mathcal{L}(Y_1, \ldots, Y_{N(t)})$ we derive the log SPRT process $u_t(\lambda_0, \mu)$. We begin from

$$\mathcal{L}(Y_1, \ldots, Y_{N(t)}) = \frac{L(Y_1, \ldots, Y_{N(t)}; (\lambda_1, \mu))}{L(Y_1, \ldots, Y_{N(t)}; (\lambda_0, \mu))} \tag{16}$$

$$= \prod_{i=1}^{N(t)} \frac{e^{-\lambda_1 \mu t}(\lambda_1 \mu t)^{Y_i}}{Y_i!} \frac{Y_i!}{e^{-\lambda_0 \mu t}(\lambda_0 \mu t)^{Y_i}} \tag{17}$$

$$= e^{-(\lambda_1 - \lambda_0)\mu t} \prod_{i=1}^{N(t)} \left(\frac{\lambda_1}{\lambda_0}\right)^{Y_i} \tag{18}$$

$$= e^{-(\lambda_1 - \lambda_0)\mu t} \left(\frac{\lambda_1}{\lambda_0}\right)^{Q(t)} \tag{19}$$

$$\ell(Y_1, \ldots, Y_{N(t)}) = Q(t) \log \left[\frac{\lambda_1}{\lambda_0}\right] - (\lambda_1 - \lambda_0)\mu t. \tag{20}$$

Our stopping time, τ, can be concisely expressed as

$$\tau = \inf\{t \geq 0; u_t(\lambda, \mu) \notin (0, h)\}. \tag{21}$$

Appendix: Deriving the CUSUM Stochastic Process, $y_t(\lambda_0, \mu)$[8]

1. Let $Y_1, Y_2, \ldots Y_n$, be the count of defects in n independent increments.
2. Let $Y_1, \ldots, Y_{\nu-1}$, be Poisson distributed with parameter λ_0, and
3. $Y_\nu, \ldots, Y_n, n \geq \nu$ be Poisson distributed with an increased parameters λ_1.

We test the composite hypotheses:

$$H_0 : \ell(Y_1, \ldots, Y_n) < h$$
$$H_\nu : \ell(Y_\nu, \ldots, Y_n) > h; \forall (1 \leq \nu \leq n)$$

To *minimize* the delay between when the change happens, ν, and the stopping time τ. We must *maximize* the difference between $\ell(Y_1, \ldots, Y_n)$ and $\ell(Y_1, \ldots, Y_\nu)$. Before ν, $\ell(Y_1, \ldots, Y_n) = u_t$ **reaches a new minimum** upon arrival of a new Y. Therefore the log likelihood ratio statistic is

$$\max_{1 \leq \nu \leq n} [\ell(Y_1, \ldots, Y_n) - \ell(Y_1, \ldots, Y_{\nu-1})] = u_t - \min_{1 \leq \nu \leq n} (u_t). \tag{22}$$

The stopping time, τ, is the first n where $u_t - \min_{1 \leq \nu \leq n} (u_t) \geq h$.

Appendix: Demo Notebooks

1. Link to Notebook to show a Poisson 2-CUSUM Scheme
2. Link to Notebook to show a Poisson/Normal HCPP 2-CUSUM Scheme
3. Link to Notebook to demo using crobat
4. Github link to crobat, data collection library I developed

B. Ramirez and Runger G.
Quantitative techniques to evaluate process stability.
Journal of Quality Engineering, 18:53–68, 2006.

J.L. Szarka and W.H. Woodall.
On the equivalence of the bernoulli and geometric cusum charts.
Journal of Quality Technology, 44:54–62, 2012.

Coinbase LLC.
Btc to usd coinbase pro — digital asset exchange, 2020.

T. Podlaski.
Copra, 2019.

S.M. Ross.
Introduction to Probability Models.
Elsevier, 2014.

F20 series film thickness measurement instruments, 2022.

J.L Devore and K.N. Berk.
Modern Mathematical Statistics with Applications.
Springer, 2012.

D. Siegmund.
Sequential Analysis.
Springer-Verlag, 1985.

V.H. Poor and O. Hadjiliadis.
Quickest Detection.
Cambridge University Press, 2009.