

eqr275

CHANGE-POINT METHODS

Douglas M. Hawkins

School of Statistics
University of Minnesota
313 Ford Hall
224 Church Street
Minneapolis, MN 55455
dhawkins@umn.edu

Peihua Qiu

School of Statistics
University of Minnesota
301389 Ford Hall
224 Church Street
Minneapolis, MN 55455
qiu@stat.umn.edu

K. D. Zamba

College of Public Health
Department of Biostatistics
The University of Iowa
200 Hawkins Dr C227 GH
Iowa City, IA 52242
gideon-zamba@uiowa.edu

Abstract

Change-point methodologies applied to statistical process control are predicated on the possibility that a special cause induces a shift from an in-control statistical model to an out-of-control statistical model, and so are particularly attractive for persistent special causes. Along with indications of a loss of control, they provide estimates of when the shift occurred, and (if needed) of the before- and after-shift process parameters. A less obvious advantage is that some change-point proposals allow for the near-universal situation that the in-control distribution of process readings is not known exactly. This feature largely removes the need for extensive Phase I calibration studies, and allows Phase II production use to start early. In addition to the stand-alone use of change-point methodologies for both signalling and diagnosing the effects of special causes, they have been proposed as tools for following up signals given by other charting methods, when their likelihood properties lead to good estimators of the time of occurrence and effect of the special cause.

Keywords: Change-point, LRT, GLR, Phase I, Phase II, SPC.

1 INTRODUCTION

Statistical process control methods are intended to detect departures from an in-control state which are the result of special cause variability. Two broad classes of departures are isolated or intermittent ones, in which the special cause appears and then disappears even in the absence of some corrective action, and sustained or persistent ones that, once introduced, tend to continue until corrective action remedies them. Persistent special causes further subdivide into those whose effects are erratic, and those that shift the process from its in-control state to a different stable out-of-control state. These latter special causes provide the motivation for the change-point formulation.

2 NOTATION AND APPROACHES

From the statistical viewpoint, the in-control and out-of-control states are distinguished by the distribution of the process measurement stream. Write $X_1, X_2, \dots, X_n, \dots$ for the sequence of readings (which could be uni- or multivariate). A convenient starting point for discussion is that, while in control, the readings are independent and follow some statistical distribution $f(x, \theta)$ with the parameter vector θ having its in-control value θ_0 . Under a persistent change at some time τ , the readings follow this distribution up to time τ , after which their behavior changes to something else. Under the conventional change-point model, this changed behavior has the same distribution with the parameter θ taking the new value θ_1 . If the change-point τ were known, we would have a standard two-sample comparison. Sometimes this is the case—we know that something happened at time τ (for example a plant switched suppliers at time τ) and want to test whether it had a visible effect on the process. More commonly, we do not know that changes have intervened until we detect and diagnose their effects. In line with this situation, we will assume τ is unknown.

If both θ_0 and θ_1 are known, then the SPC diagnostic of choice is generally a cumulative sum (cusum) chart, with its minimum time to detection property. The change-point formulation is therefore of interest in settings where at least one of these parameters is unknown. As it is rare for the out-of-control distribution to be known while the in-control is not, we ignore this possibility,

and consider the situations

- θ_0 known, θ_1 unknown; and
- Both θ_0 and θ_1 unknown.

The approach to testing for a change-point is the same; the log likelihood of the sequence X_1, \dots, X_n under the change-point model is the sum of two terms: one for the in-control and one for the out-of-control portions of the sequence:

$$\log L(\theta_0, \theta_1, \tau) = A(\theta_0, \tau) + B(\theta_1, \tau, n),$$

where $A(\theta_0, \tau) = \sum_{i=1}^{\tau} \log f(X_i, \theta_0)$ is the in-control portion, and $B(\theta_1, \tau, n) = \sum_{i=\tau+1}^n \log f(X_i, \theta_1)$ is the out-of-control. If any of the parameters are unknown, then they are estimated by maximizing the likelihood. So if, for example, θ_1 is unknown as it commonly is, then it is replaced by the value $\hat{\theta}_1$ maximizing the $B(\theta_1, \tau, n)$ term, and similarly for θ_0 . The change-point τ is estimated as the value k maximizing the sum $A(\theta_0, k) + B(\hat{\theta}_1, k, n)$ if θ_0 is known, and $A(\hat{\theta}_0, k) + B(\hat{\theta}_1, k, n)$ if it is not. If there is no change-point, then the log likelihood is $A(\theta_0, n)$. Thus the conventional generalized likelihood ratio (GLR) test for the presence of a change-point is

$2[\log L(\hat{\theta}_0, \hat{\theta}_1, \hat{\tau}) - A(\hat{\theta}_0, n)]$ if θ_0 is unknown, and

$2[\log L(\theta_0, \hat{\theta}_1, \hat{\tau}) - A(\theta_0, n)]$ if it is known.

As with many other likelihood ratio methods, this test can often be re-expressed in different but functionally equivalent forms.

3 NORMALLY DISTRIBUTED DATA

The easiest and most familiar example is normally distributed data. Switching to a more standard notation, suppose that the pre-shift in-control distribution of the data is $N(\mu_0, \sigma_0^2)$, and the post-shift out-of-control distribution is $N(\mu_1, \sigma_1^2)$. It is most commonly assumed that the mean changes following the shift, i.e. $\mu_0 \neq \mu_1$. One family of change-point formulations, the homoscedastic models, assumes that $\sigma_0 = \sigma_1 = \sigma$ say, and the other, the heteroscedastic model, allows $\sigma_0 \neq \sigma_1$.

Write $\bar{X}_{ik} = \sum_{j=i+1}^k X_j / (k - i)$ for the mean of the sequence X_{i+1}, \dots, X_k and

$V_{ik} = \sum_{j=i+1}^k (X_j - \bar{X}_{ik})^2$ for the sum of squared deviations of these observations about their mean.

Then splitting at a trial value $\tau = k$, gives maximum likelihood estimators

$\hat{\mu}_0 = \bar{X}_{0k}$, $\hat{\mu}_1 = \bar{X}_{kn}$ for the means, and for the unequal variance setting, $\hat{\sigma}_0^2 = V_{0k}/(k - 1)$ and $\hat{\sigma}_1^2 =$

$V_{kn}/(n - k - 1)$ for the separate variances, or, in the homoscedastic case, $\hat{\sigma}^2 = (V_{0k} + V_{kn})/(n - 2)$.

These variance estimators are not, strictly speaking, MLEs, as they have been scaled by their degrees of freedom, and not by the sample sizes on which they are based, but this situation is standard.

3.1 HOMOSCEDASTIC CASE

In the homoscedastic case, the estimator of the change-point τ is the value of k maximizing, in absolute value,

$$U_{kn} = \sqrt{\frac{k(n-k)}{n}} (\bar{X}_{0k} - \bar{X}_{kn}),$$

or equivalently, minimizing $V_{0k} + V_{kn}$

Write $U_{\max,n}$ for this maximized absolute value. The test statistic for a change-point at an unknown time τ is then $U_{\max,n}/\sigma$ if σ is somehow known, and in the more common situation that it is not known, is $T_{\max,n} = U_{\max,n}/\hat{\sigma}$. This is just the absolute maximum across k of the two-sample t test comparing the ‘samples’ X_1, \dots, X_k and X_{k+1}, \dots, X_n .

The test can also be expressed as the maximum of the F ratio in an analysis of variance comparing these two samples as

$$F_{\max,n} = \max_k \left[\frac{V_{0n} - V_{0k} - V_{kn}}{V_{0k} + V_{kn}} (n - k - 2) \right] = T_{\max,n}^2.$$

In the setting where μ_0 is already known exactly, U_{kn} is redefined as $U_{kn} = \sqrt{n-k}(\mu_0 - \bar{X}_{kn})$ and there is a corresponding minor adjustment in the pooled standard deviation if σ has to be estimated, but the broad thrust of the procedure is unchanged. See related publications such as [5] [6] [9] and [12].

3.2 HETEROSCEDASTIC CASE

There are two interesting change-point tests to perform in the heteroscedastic setting. One is to test for a change in the variance without regard to whether or not the mean changed when the special cause intervened. The other is to test the possibility of a change in the mean, the variance, or both. Likelihood ratio test statistics for these possibilities start out with:

Test variance only: [4]

$$G_{kn} = [(k-1) \log(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}) + (n-k-1) \log(\frac{\hat{\sigma}^2}{\hat{\sigma}_1^2})]/B,$$

with

$$B = 1 + [(k-1)^{-1} + (n-k-1)^{-1} - (n-2)^{-1}]/3$$

Test mean and/or variance shift: [12]

$$G_{kn} = [k \log\{\frac{kV_{0n}}{nV_{0k}}\} + (n-k) \log\{\frac{(n-k)V_{0n}}{nV_{kn}}\}]/B,$$

with

$$B = 1 + \frac{11}{12}[k^{-1} + (n-k)^{-1} - n^{-1}] + [k^{-2} + (n-k)^{-2} - n^{-2}].$$

The constants B are Bartlett correction factors intended to improve the approximation of the distribution of these statistics to chi-squared distributions with degrees of freedom 1 and 2 respectively for fixed k and n . As with the shift-in-mean problem, the unknown time of change τ is estimated by the k maximizing the test statistic.

4 THEORETICAL RESULTS

For a given k , U_{kn}/σ follows a standard normal distribution, and T_{kn} follows a t distribution with $n-2$ degrees of freedom. Maximization of either statistic over k however takes the distributions out of the realm of these familiar standard distributions. There is a vast literature on the fixed-sample homoscedastic mean change-point problem where the sample size n is fixed (see for example [1] and [2]) where the technical difficulties are discussed in detail. This situation of a fixed sample size is

relevant in a quality setting to the analysis of Phase I data. Specific applications to SPC include [11].

5 PHASE II SPC

In Phase II SPC, the sample size is not fixed. Rather while the process is thought to be in control, additional readings are added to the record and incorporated into the control charts. As each new reading is obtained, a decision has to be made of whether the process still appears to be in control. If so, then the reading is incorporated into the history and the process is allowed to continue. If however the reading gives rise to a suspicion that the process is now out of control, then the process is stopped for diagnosis and remedial action.

Returning to the generic change-point formulation, a purely-change-point-driven procedure can be described as:

- When each new observation accrues, calculate

$$C_n = \max_k 2[\log L(\hat{\theta}_0, \hat{\theta}_1, k) - A(\hat{\theta}_0, n)]$$

- If $C_n < c_n$ then conclude that there is no evidence of a shift.
- If $C_n > c_n$ then conclude that there is evidence of a departure from control.

The sequence c_n is a sequence of control limits chosen to control the false-alarm behavior of the scheme. Finding this sequence c_n does not seem to be amenable to theoretical analysis, and actual implementations of the change-point method have used simulation to find suitable values.

A particular attraction of the change-point method is its post-signal diagnostic capability. Under the model that the process did indeed have a persistent step change from the in-control to the out-of-control parameter value, the k value maximizing the likelihood ratio gives an estimate of τ , the last reading before the shift, and the corresponding $\hat{\theta}_1$ provides a maximum likelihood estimate of the out-of-control parameter value. While maximum likelihood, however, these estimates do not necessarily have the usual ‘good’ properties of maximum likelihood estimators such as consistency.

6 THE CHANGE-POINT METHOD AS FOLLOW-UP

This sketch is of the use of the change-point method as a self-contained Phase II charting methodology. This is not the only way it can be used. For example some other method such as a Shewhart, cusum or EWMA chart may be used to detect departure from control, and following this signal, the change-point calculations may be done in an essentially Phase I fashion to estimate the time of the change and the post-change parameters. Relevant articles are [8] [10] and [13].

In this use, the conventional distributional results of the pure change-point theory are largely inapplicable. Test statistic distributions, for example, assume that the data follow some specified (for example common normal) distribution; these distributions do not apply when we condition on the fact that the data look suspicious to some other charting method.

7 EXISTING CHANGE-POINT FORMULATIONS

The change-point framework is general; specific existing applications are:

- Normal data, where the mean may undergo a step change [5].
- Normal data, where the variance may undergo a step change [4].
- Normal data, where the mean and/or variance may undergo a step change [3].
- Multivariate normal data, where the mean vector may undergo a step change [14].
- Regression data, where the dependence of a process variable on some covariate may undergo

a step change [7].

The framework is however more general than a menu of existing implementations might suggest. There is nothing to prevent, for example, the in-control distribution being, not one of independent identically distributed readings, but say a time series, modeled by an autoregressive integrated moving average scheme whose coefficients can change when the process goes out of control. This model involves additional computational, but not conceptual, complexity.

8 MULTIPLE CHANGE-POINTS

The discussion so far has assumed that the process history shows a maximum of one change-point. This raises the question of whether attention to multiple change-points is needed. There is a case for arguing that it is not; that when the process goes out of control, the process owner should diagnose the situation as soon as the control scheme provides an alert. Following this, the data identified as seeming out of control should be excised from the in-control record, and when charting starts up again, the change-point calculations restarted from the last in-control point, and new process readings added sequentially.

9 COMPARISON WITH OTHER METHODS

As noted earlier, SPC settings divide into those involving transient and persistent special causes. The latter may give rise to either erratic data or some shifted statistical distribution. A further distinction is between settings where the in-control statistical distribution is known exactly (perhaps as a result of an earlier large Phase I study) and those in which the in-control parameters are not known exactly.

The standard SPC methods of Shewhart, cusum and EWMA charting assume that the in-control parameters are known exactly. The more common reality is that they are estimated from finite Phase I samples, and recent research has shown that these Phase I samples need to be very large if the known-parameter run behavior is to be realistic.

The difficulties caused by these imperfectly-known in-control parameters can logically be accommodated in two ways - by adapting the control limits or interpretive rules of the standard charts; and by using methodologies that accommodate unknown parameters explicitly. The self-starting approach is an example of the latter, as is the unknown-parameter change-point formulation. They differ in that the self-starting method makes no assumption about the out-of-control distribution of the data whereas the change-point formulation does. Logically, if the change is persistent and leads to stable behavior, then change-point methods should be the methods of choice; if the change

is persistent but leads to erratic behavior, then the unmet assumption of stable behavior should make change-point formulations less effective.

10 EXAMPLE: GOLD MINE SAMPLING QC

This example fits the more general scenario where changes in both mean and/or variance are important and can hinder the performance of a process. We refer the reader to [3] for a fuller discussion and details on this example. The objective is to monitor gold mine samplers. Samplers in gold mines chip out rock samples of the “face” where the ore is being extracted and submit them for chemical assay to measure their gold content. As a quality check, supervisors cut out fresh samples at some of the locations already sampled. This gives rise to pairs of samples and of gold content – one by the original sampler and one by the supervisor, an experienced sampler. The log of the ratio of the gold content of the sampler to that of the supervisor has an approximately normal distribution. This distribution should have a zero mean (else the sampler is biased), and a small variance (else the sampler is erratic). Figure 1 shows a sequence of such log ratios for a junior sampler.

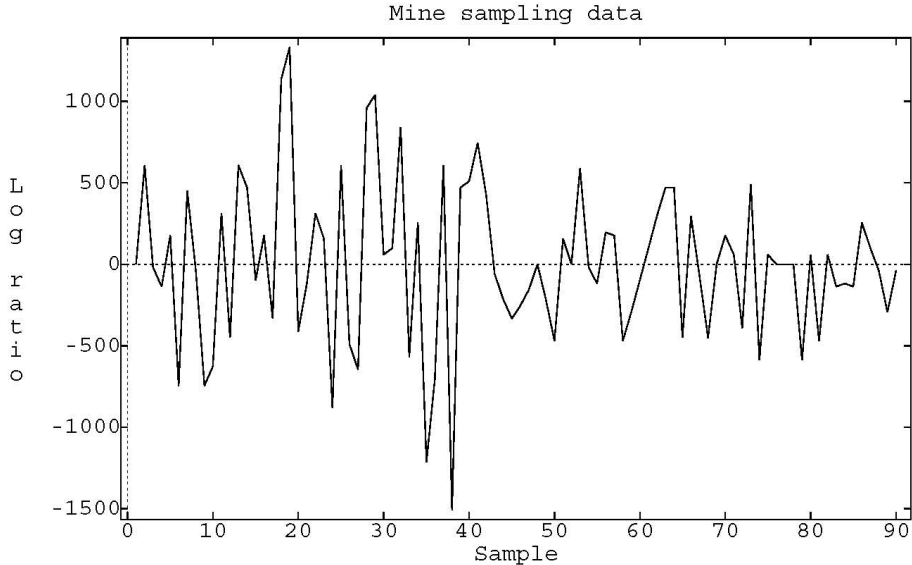


Figure 1: Sample vs log ratio sampler over supervisor

Change points in both mean and variance are possible and important. A change in mean leads to a bias in mine valuation. An increase in variance may warn of loss of motivation to sample carefully, while a decrease in variance indicates a highly desirable improvement in measuring quality, perhaps as a result of learning skills. We will address these questions using the more general setting of change- point formulation to look for changes in this sampler’s output. Visually, Figure 1 suggests that the variability in the portion up to around observation 40 is higher than that for the remainder of the sequence. The mean however appears to center on zero for the whole sequence. Figure 2 gives the chart of the GLR along with the control limit $c_{n,0.002}$, where $c_{n,0.002}$ is chosen to yield an in-control ARL of 500 for normal data. The chart first crosses the control limit at reading 70,

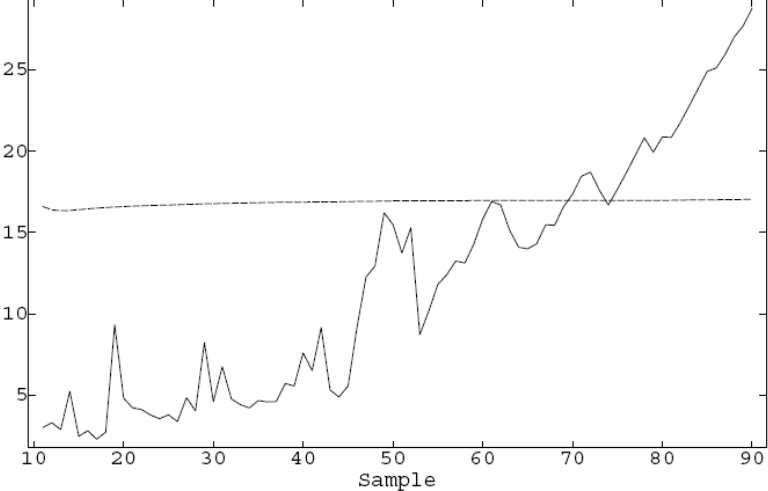


Figure 2: C_n in solid line and $c_{0.002,n}$ in dashed line

and continues upward except for a brief dip below at reading 74. Right from the first hint of a change point, the GLR gives the left-segment estimates $\hat{\tau} = 42, \hat{\mu}_1 = 0.053, \hat{\sigma}_1 = 0.42$. Table 1 lists the summary statistics of the right segment parameters, the GLR, the t and the F test statistics for identity of mean and of variance, and the base-10 log of their normal-distribution P values. These confirm that there is no indication of a shift in mean, but that the standard deviation is substantially smaller after the change point than before. By the time of the first signal at sample

number 70, the estimated post-change standard deviation has stabilized close to its ultimate value of 0.283 – about two-thirds of its value in the first portion of the series. This reduction of variance is of substantial value to the mine. It means that each sample now produced by this junior sampler is more informative than two of his previous samples were. This has obvious implications for the improvement in mine valuation and selection of which ore to extract.

Table 1. The Gold Mine Summary Statistics

i	$\bar{X}_{\hat{\tau},i}$	$sd_{\hat{\tau},i}$	C_n	t	$\log P$	F	$\log P$
69	-0.035	0.297	16.60	0.85	-0.40	4.74	-4.39
70	-0.027	0.294	17.36	0.79	-0.36	4.83	-4.59
71	-0.024	0.289	18.47	0.77	-0.35	4.99	-4.85
72	-0.037	0.292	18.72	0.88	-0.42	4.90	-4.88
73	-0.020	0.303	17.60	0.72	-0.33	4.57	-4.68
74	-0.037	0.314	16.70	0.87	-0.41	4.24	-4.44
75	-0.034	0.310	17.66	0.86	-0.40	4.36	-4.66
76	-0.033	0.305	18.70	0.85	-0.40	4.50	-4.91
77	-0.032	0.301	19.76	0.85	-0.40	4.63	-5.15
78	-0.032	0.296	20.84	0.85	-0.40	4.77	-5.40
79	-0.047	0.306	19.94	0.98	-0.48	4.47	-5.16
80	-0.044	0.302	20.87	0.96	-0.47	4.58	-5.38
81	-0.055	0.306	20.86	1.06	-0.53	4.47	-5.34
82	-0.052	0.303	21.76	1.03	-0.52	4.57	-5.55
83	-0.054	0.299	22.79	1.06	-0.53	4.67	-5.78
84	-0.056	0.296	23.85	1.08	-0.54	4.79	-6.01
85	-0.057	0.292	24.91	1.10	-0.56	4.89	-6.24
86	-0.050	0.293	25.11	1.03	-0.52	4.88	-6.32
87	-0.047	0.290	25.94	1.01	-0.50	4.97	-6.53
88	-0.047	0.287	27.01	1.01	-0.50	5.08	-6.77
89	-0.052	0.286	27.71	1.06	-0.53	5.11	-6.90
90	-0.052	0.283	28.79	1.06	-0.53	5.22	-7.15

11 RELATED ARTICLES

eqr245, eqr246, eqr247, eqr248, eqr249, eqr251, eqr253.

REFERENCES

- [1] Carlstein, E , Mueller, H G and Siegmund, D. *Change-Point Problems*. IMS Lecture Notes **23**, Springer, 1994.
- [2] Csörgö, M , and Horváth, L. *Limit Theorems in Change-Point Analysis*. John Wiley & Sons: New York, Chichester, 1997
- [3] Hawkins, D M and Zamba, K D. Statistical Process Control for Shifts in Mean or Variance using a Change Point Formulation. *Technometrics* **47**, 164–173, 2005.
- [4] Hawkins, D M and Zamba, K D. A Change Point Model for a Shift in Variance. *Journal of Quality Technology* **37**, 21–31, 2005.
- [5] Hawkins, D M , Qiu, P and Kang, C W. The Change-point Model for Statistical Process Control. *Journal of Quality Technology* **35**, 355–365, 2003.
- [6] Lai, T L. Sequential Analysis: Some Classical Problems and New Challenges. *Statistica Sinica* **11**, 303–350, 2001.
- [7] Mahmoud, M A , Parker, P A , Woodall, W H, and Hawkins, D M. A Change Point Method for Linear Profile Data. To appear in *Quality and Reliability Engineering International*, 2006.
- [8] Pignatiello, J J Jr, and Samuel, T R. Estimation of the Change Point of a Normal Process Mean in SPC Applications. *Journal of Quality Technology* **33**, 82–95, 2001.
- [9] Siegmund,D and Venkatraman,E S. Using the Generalized likelihood Ratio Statistics for Sequential Detection of a Change point. *Annals of Statistics* **23**, 255–271, 1994.
- [10] Srivastava, M S, Wu, Yanhong. Quasi-Stationary Biases of Change Point and Change Magnitude Estimation After Sequential CUSUM Test. *Sequential Analysis* **18**, 203–216, 1999.
- [11] Sullivan, J H , Woodall, W H. A Control Chart for Preliminary Analysis of Individual Observations. *Journal of Quality Technology* **28**, 265–278, 1996.
- [12] Worsley, K J. On the Likelihood Ratio Test for a Shift in Location of Normal Populations. *Journal of American Statistical Association* **74**, 365–367, 1979.
- [13] Wu, Yanhong. Inference for Change-Point and Post-Change Means After a CUSUM Test. *Lecture Notes in Statistics* **180**, Springer: New York, 2005.
- [14] Zamba, K D, and Hawkins, D M. A Multivariate Change Point Model for Statistical Process Control. *Technometrics* **48**, 4 ,539-549, 2006.