Controlling Correlated Processes with Binomial Marginals

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics

This talk is based on the paper

Weiß, C. H.:
Controlling Correlated Processes with Binomial Marginals.
Preprint 277, Mathematische Institute der
Julius-Maximilians-Universität Würzburg, 2007.
See conference CD-ROM.

All references mentioned in this talk correspond to the references in this article.

One year ago
in Wroclaw . . .

INAR(1) model for processes of counts:

Let $(\epsilon_t)_{\mathbb{N}}$ be i.i.d. process with range $\mathbb{N}_0$, let $\alpha \in [0;1]$. An INAR(1) process $(N_t)_{\mathbb{N}_0}$ follows the recursion

$$N_t \;=\; \alpha \circ N_{t-1} \;+\; \epsilon_t, \qquad\qquad t \geq 1.$$

McKenzie (1985), Al-Osh & Alzaid (1987, 1988)

Binomial thinning, due to Steutel & van Harn (1979):

N discrete random variable with range $\{0,\ldots,n\}$ or $\mathbb{N}_0$.
Define random variable

$$\alpha \circ N \;:=\; \sum_{i=1}^{N} X_i,$$

where X_i are independent Bernoulli trials, $B(1,\alpha)$, also independent of N $\rightarrow$ *counting series*.

We say: $\alpha \circ N$ arises from N by *binomial thinning*
'$\circ$' is called *binomial thinning operator*.

Interpretation of $\alpha \circ N$:

- Population of size N at a certain time t.

- Later at time $t+1$: population shrinked, because some individuals died.

- Assume that individuals die independently of each other with probability $1-\alpha$

$\Rightarrow$ *Number of survivors* is given by $\alpha \circ N$.

Motivation

The INAR(1) process . . .

- is easy to interpret,
- is well-suited for many popular count distributions: Poisson, negative binomial, generalized Poisson,
- applies well to typical tasks of SQC,
- can be controlled efficiently, . . .

For details, see

Weiß, C.H.: *Controlling correlated processes of Poisson counts*. QREI 23(6), 2007.

. . . but by definition

$$N_t \;=\; \alpha \circ N_{t-1} \;+\; \epsilon_t, \qquad t \geq 1.$$

of the INAR(1) process, the INAR(1) model can be applied to processes of counts with the infinite range $\mathbb{N}_0$ only!

The
Binomial AR(1) Model

Definition & Interpretation

Let $n \in \mathbb{N}$, $p \in (0;1)$ and $\rho \in [\max\left(-\frac{p}{1-p}, -\frac{1-p}{p}\right); 1]$.

Define $\beta := p \cdot (1-\rho)$ and $\alpha := \beta + \rho$.

The process $(X_t)_{\mathbb{N}_0}$ with

$$X_t \;=\; \alpha \circ X_{t-1} \;+\; \beta \circ (n - X_{t-1}), \quad t \geq 1, \qquad X_0 \sim B(n,p),$$

where all thinnings are performed independently of each other, and the thinnings at time t are independent of $(X_s)_{s<t}$, is called a **binomial AR(1) process**.

McKenzie (1985)

Interpretation of $X_t = \alpha \circ X_{t-1} + \beta \circ (n - X_{t-1})$:

System of n independent units, either in state 1 or state 0.

X_{t-1}: number of units in state 1 at time $t-1$.

$\alpha \circ X_{t-1}$: number of units still in state 1 at time t, with individual transition probability α.

$\beta \circ (n - X_{t-1})$: number of units, which moved from state 0 to state 1 at time t, with individual transition probability β.

Examples: $X_t \;=\; \alpha \circ X_{t-1} \;+\; \beta \circ (n - X_{t-1})$

- Computer pool with n machines, either occupied (state 1) or not (state 0). Here, X_t is number of machines occupied at time t, consisting of machines occupied before, and machines newly occupied.
- Hotel rooms in certain hotel being occupied at day t . . .
- Clerks in a counter room serving a customer . . .
- Telephones in a call centre being occupied, etc.

The Binomial AR(1) Model

Properties

Let $(X_t)_{\mathbb{N}_0}$ be binomial AR(1) process.

- $(X_t)_{\mathbb{N}_0}$ is a stationary Markov chain with marginal distribution $B(n,p)$

- transition probabilities

$$p_{k|l} := P(X_t = k \mid X_{t-1} = l) \; = \; \sum_{m=\max(0,k+l-n)}^{\min(k,l)} \binom{l}{m} \binom{n-l}{k-m} \, \alpha^m (1-\alpha)^{l-m} \; \beta^{k-m} (1-\beta)^{n-l+m-k}.$$

(. . .)

- autocorrelation function

$$\rho(k) := Corr[X_t, X_{t-k}] \;=\; \rho^k, \qquad k \geq 0$$

Remark: The autocorrelation function of $(X_t)_{\mathbb{N}_0}$ is that of a usual AR(1) process.
The occurrence of negative autocorrelation is equivalent to $\alpha < \beta$, i. e., it is more likely to reach state 1 from state 0 than from state 1.

(...)

- conditional moments:

$$E[X_t \mid X_{t-1}] \;=\; \rho \cdot X_{t-1} \;+\; n\beta, \qquad \text{and}$$
$$V[X_t \mid X_{t-1}] \;=\; \rho(1-\rho)(1-2p) \cdot X_{t-1} \;+\; n\beta(1-\beta).$$

- $r_i(k) := P(X_t = \ldots = X_{t+k-1} = i \mid X_{t-1} \neq i, X_t = i)$, $i = 0, \ldots, n$ and $k \in \mathbb{N}$: Conditional probability that i-run, starting at time t, is of length at least k, conditioned on event that an i-run started at time t.

$$r_i(k) \;=\; p_{i|i}^{k-1}, \qquad k \geq 1, \qquad \text{and} \qquad \mu_i \;=\; 1/(1-p_{i|i}).$$

The
Binomial AR(1) Model

Model Building

Model identification:

- Autocorrelation of AR(1) type: empirical partial autocorrelations $\hat{\rho}_{\mathrm{p}}(k)$ should be about 0 for $k > 1$.

- Histogram and Pearson's χ^2-test. But both very sensible to autocorrelation:

$$\mathrm{X}_g^2 \ := \ \sum_{i=0}^{n} \frac{(N_i - Tp_i)^2}{Tp_i} \quad \xrightarrow{D} \quad \sum_{j=1}^{n} \frac{1+\rho^j}{1-\rho^j} \cdot Z_j^2,$$

where $Z_1, \ldots, Z_n$ i.i.d. $N(0,1)$.

Proof: See Weiß (2007).

Model Estimation:

- *Yule-Walker approach*: Estimate p by mean $\frac{1}{n(T+1)} \cdot \sum_{t=0}^{T} X_t$, and ρ by first order empirical autocorrelation.

- *ML estimates*: Likelihood function determined easily, since process is simple Markov chain with $n+1$ states.

- *Conditional least squares (CLS) approach*: Since $E[X_t \mid X_{t-1}] = \rho \cdot X_{t-1} + np(1-\rho)$, minimize

$$CSS(p,\rho) \;:=\; \sum_{t=1}^{T} \big(X_t - \rho \cdot X_{t-1} - np(1-\rho)\big)^2.$$

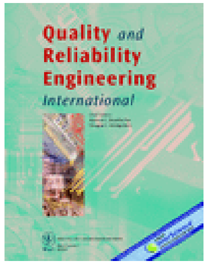
The Binomial AR(1) Model

Control Schemes

Binomial AR(1) Model

Idea:

Based on above properties of binomial AR(1) process, adapt the control schemes for Poisson INAR(1) processes, developped by



Weiß, C.H.: *Controlling correlated processes of Poisson counts*. QREI 23(6), 2007.

np-Chart for Binomial AR(1):

Realized values of $(X_t)_{\mathbb{N}}$ plotted on chart with

$$\begin{aligned} UCL &= np_0 + 3\sqrt{np_0(1-p_0)}, \\ \text{Center} &= np_0, \\ LCL &= \max\left\{0, np_0 - 3\sqrt{np_0(1-p_0)}\right\}. \end{aligned}$$

4th alternative: **Moving window** of length w.

Window sum ${C_t}^{(w)} \;:=\; N_{t-w+1} + \ldots + N_t$, with

$$E[\tfrac{1}{w}{C_t}^{(w)}] \;=\; np_0$$

and

$$V[\tfrac{1}{w}\cdot{C_t}^{(w)}] \;=\; \frac{np_0(1-p_0)}{w}\cdot\frac{1+\rho_0}{1-\rho_0}\cdot\left(1-\frac{2}{w}\cdot\frac{\rho_0}{1-\rho_0^2}\cdot(1-\rho_0^w)\right).$$

Controlling a Moving Average: Window size w, step width s. Plot statistics $\ldots, T_t{}^{(w)}, T_{t+s}{}^{(w)}, \ldots$, defined by

$$T_t^{(w)} \;:=\; \frac{1}{w}\cdot C_t^{(w)},$$

with

$$\begin{aligned} UCL &= np_0 + 3\cdot\sqrt{\tfrac{np_0(1-p_0)}{w}\cdot\tfrac{1+\rho_0}{1-\rho_0}\cdot\Big(1-\tfrac{2}{w}\cdot\tfrac{\rho_0}{1-\rho_0^2}\cdot(1-\rho_0^w)\Big)},\\ \text{Center} &= np_0,\\ LCL &= np_0 - 3\cdot\sqrt{\tfrac{np_0(1-p_0)}{w}\cdot\tfrac{1+\rho_0}{1-\rho_0}\cdot\Big(1-\tfrac{2}{w}\cdot\tfrac{\rho_0}{1-\rho_0^2}\cdot(1-\rho_0^w)\Big)}. \end{aligned}$$

If in-control autocorrelation ρ_0 very large (>0.9), then $(X_t)_{\mathbb{N}_0}$ tends to long runs of counts.

Define limits UCL_i based on probability γ, e. g., $\gamma = 0.0027$:

$$UCL_i \;=\; \left\lceil \frac{\ln \gamma}{\ln p_{i|i}} - 1 \right\rceil, \qquad i = 0, \ldots, n.$$

Monitor runs $(\boldsymbol{R}_n)_{\mathbb{N}}$, where $\boldsymbol{R}_n = (I_n, Y_n)$ represents an I_n-run of length $Y_n \geq 1$. The statistic plotted is given by

$$T_n \;:=\; \min\left(0,\; \frac{Y_n - \mu_{I_n}}{\mu_{I_n}}\right) \;+\; \max\left(0,\; \frac{Y_n - \mu_{I_n}}{UCL_{I_n} - \mu_{I_n}}\right).$$

The
Binomial AR(1) Model

A Case Study

A Case Study

- Data about log-ins and log-outs on public computerized workstations of the computer centre of the University of Würzburg.

- Workstations accessible from monday to friday during the term-time, from eight o'clock in the morning until eight o'clock in the evening.

- Any of these five working days showed a different use profile, depending on the timetable of the students.

The results presented refer to tuesdays during the term-time, beginning with May 3rd, 2005. To guarantee data as homogeneous as possible, we will restrict ourselves on the term-time between ten o'clock in the morning and half past five in the evening.

The aim of the analysis was to identify and model an in-control using profile of the workstations, and to construct control procedures based on this in-control model to identify unusual days.

In the following, we concentrate on $n = 15$ workstations, located together and fully operative during observation time.

The total count of log-ins on these 15 workstations was computed for each second.
To reduce amount of data, only counts $X_{i,t}$ at beginning of t^{th} minute on day i considered.

So for each day i in the observation period, a time series with 451 counts was available.

If it is equally probable to log in to any of the workstations observed, the marginals of the count sequences may follow a binomial distribution.

Furthermore, the occupied workstations at time t consist of workstations, which have also been occupied at time $t-1$ before, and workstations where a user logged in during the last minute.

So a binomial AR(1) model may be suitable to describe the data.

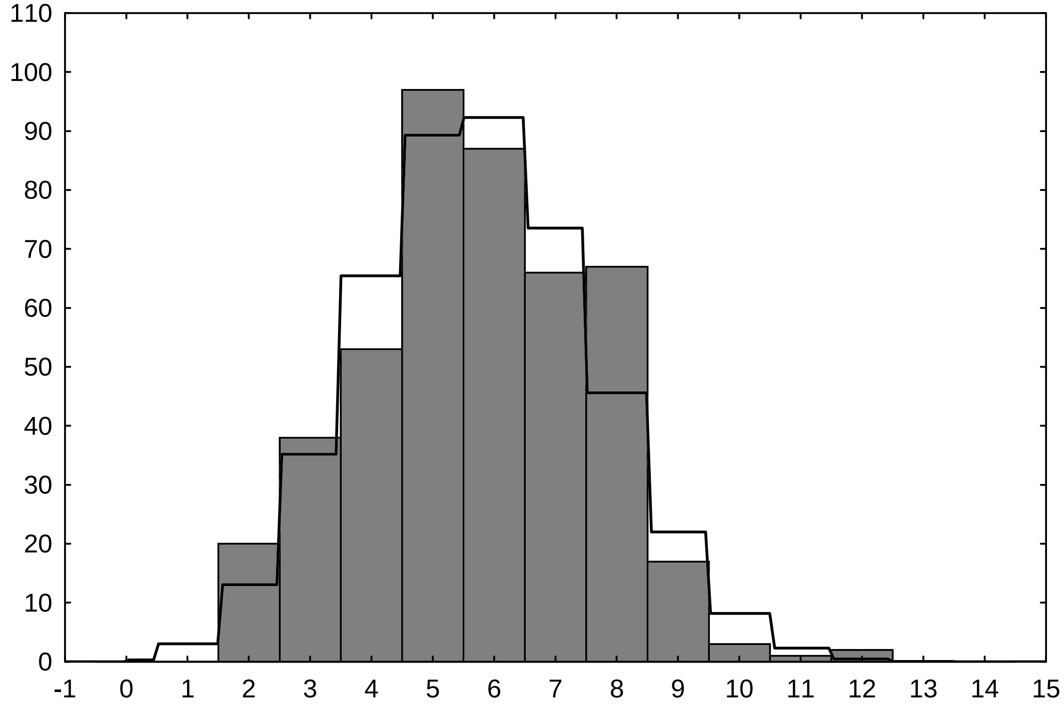
May 3rd, 2005:

The empirical mean of the data, divided by $n = 15$, equals 0.38271. So the mean probability p of a workstation being occupied is estimated by about 38 %.

Enormous extend of autocorrelation $\hat{\rho}(k)$ observed:

k	1	2	3	4	5	6	
$\hat{\rho}(k)$	0.963	0.920	0.881	0.839	0.796	0.762	...
$\hat{\rho}_{\mathrm{p}}(k)$	0.963	-0.091	0.024	-0.058	-0.031	0.096	...

A Case Study



Assumption of binomial distributed marginals is reasonable, theoretical density: $B(15, 0.38)$ distribution. Deviations caused by enormous extend of autocorrelation.

Model estimation:

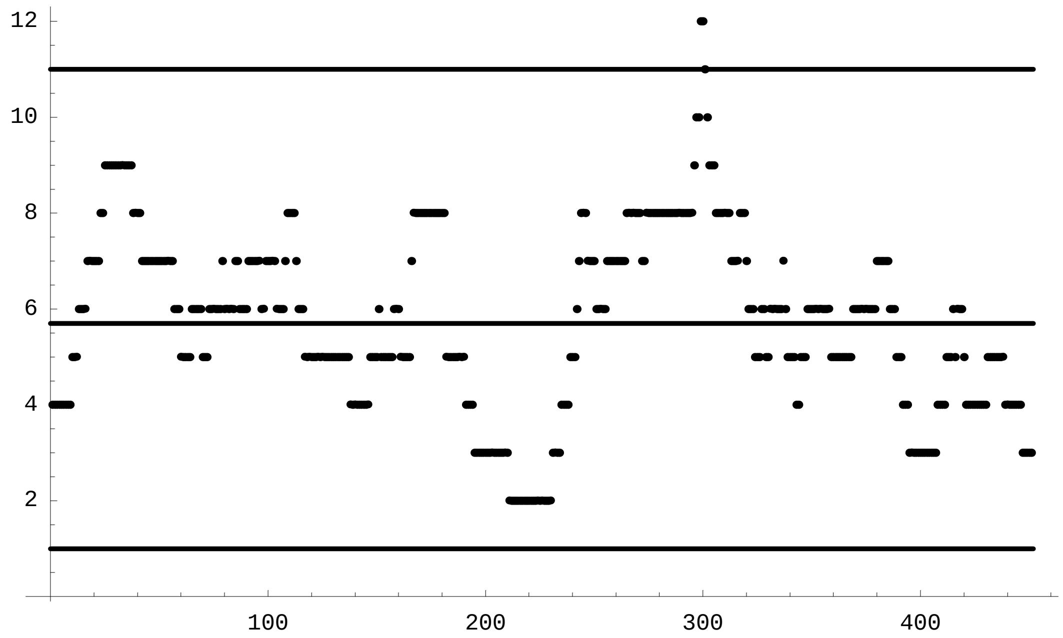
Yule-Walker estimates of p_0 and ρ_0: 0.38271 and 0.962792.

CLS estimates of p_0 and ρ_0: 0.378308 and 0.969168.

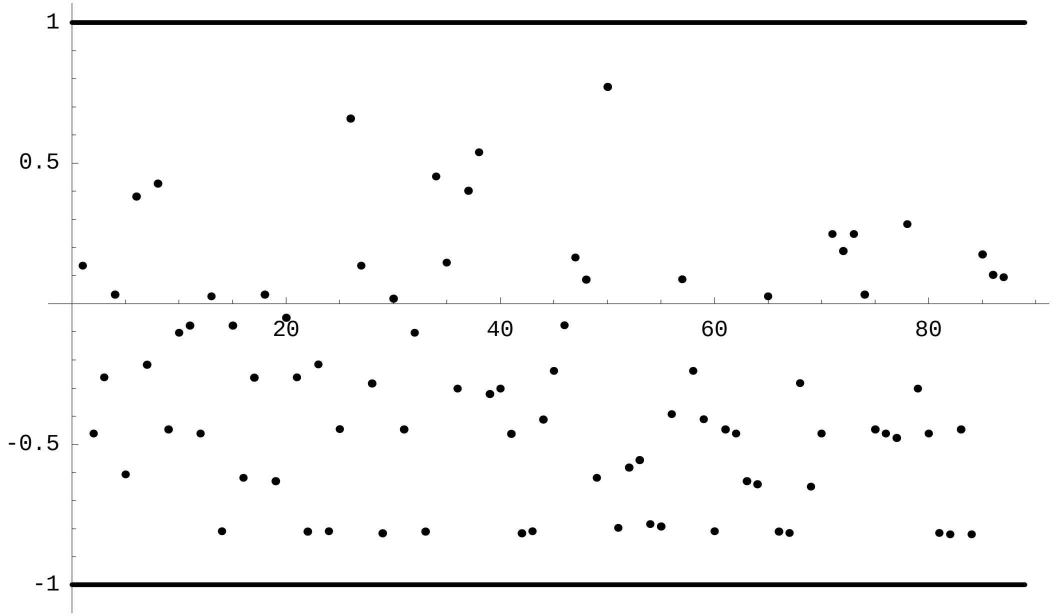
ML estimates of p_0 and ρ_0: 0.364926 and 0.968219.

In-control model: The sequences $(X_{i,t})_{t=0,\ldots,450}$ are assumed to follow a binomial AR(1) model with parameters $p_0 = 0.38$ and $\rho_0 = 0.97$ in the state of control.

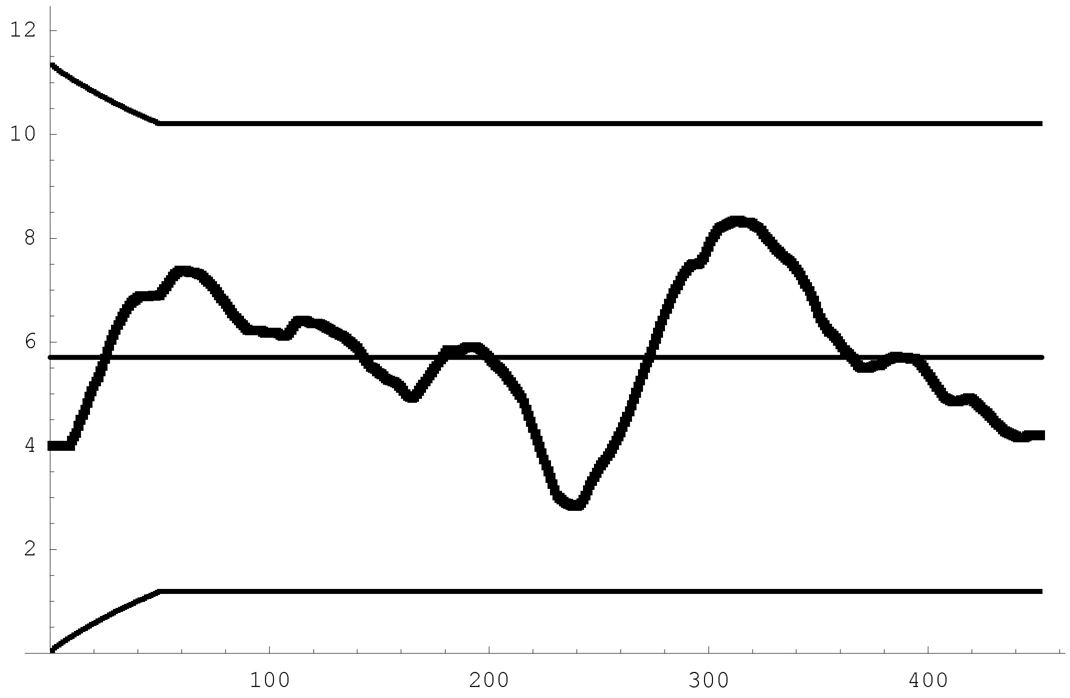
np chart of log-in counts on May, 3[rd]:



Runs chart of log-in counts on May, 3rd:



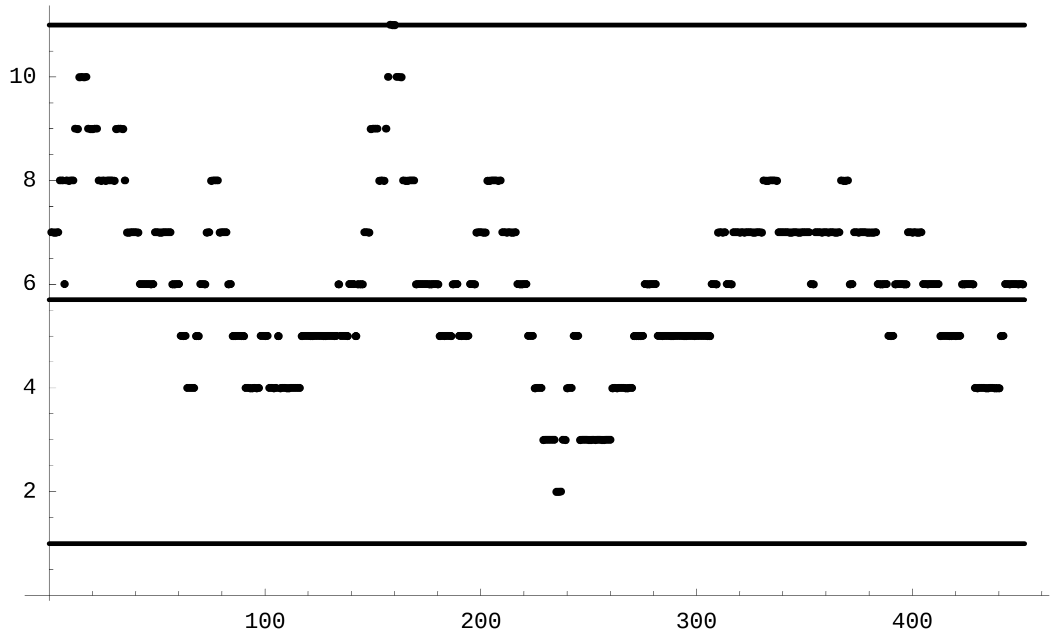
Moving-average chart ($w = 50$) on May, 3rd:



May 10th, 2005:

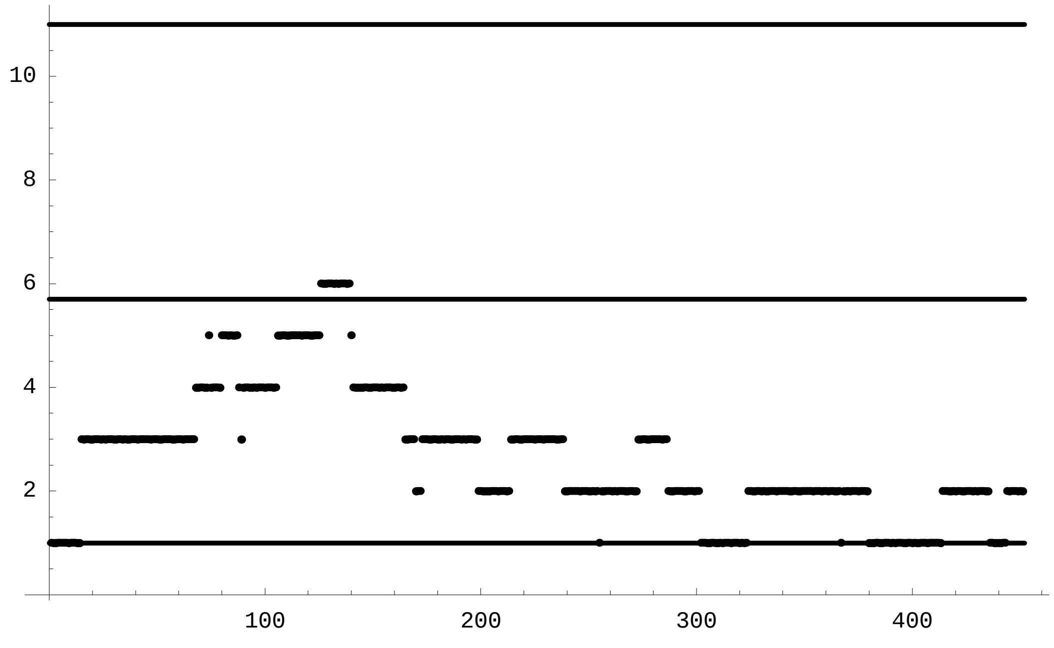
None of charts signals alarm, process seems in control.

np chart, for instance:

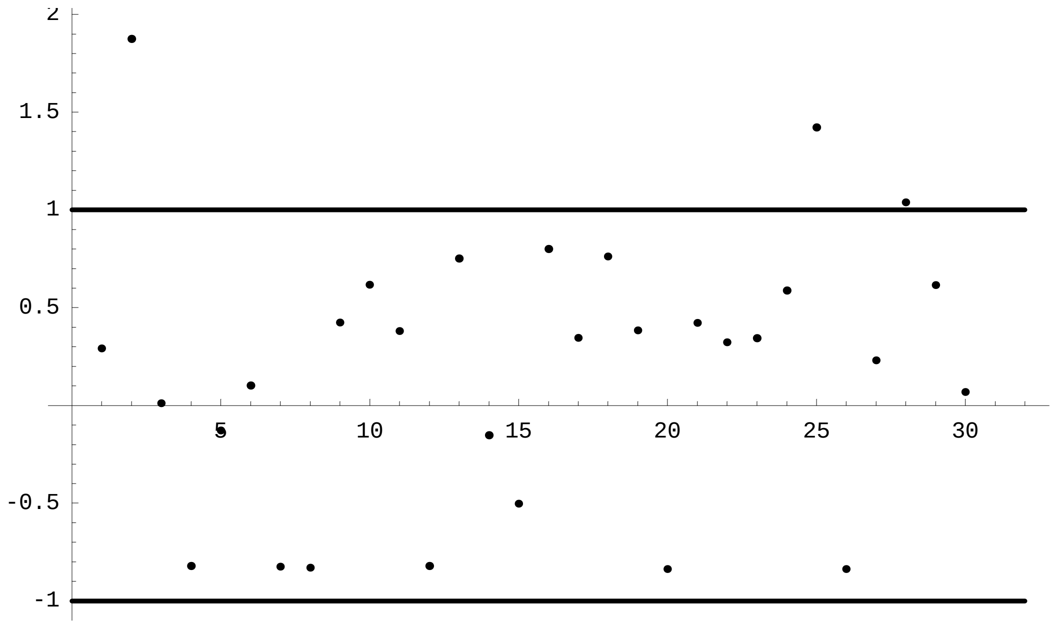


May 17th, 2005:

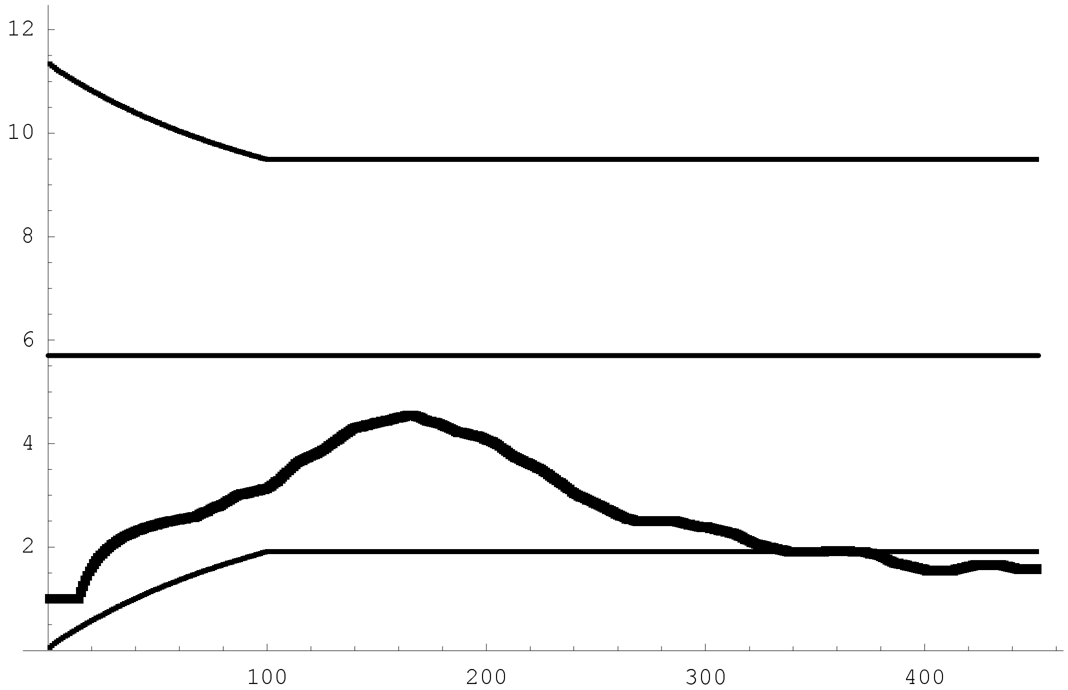
np chart of log-in counts on May, 17th:



Runs chart of log-in counts on May, 17th:

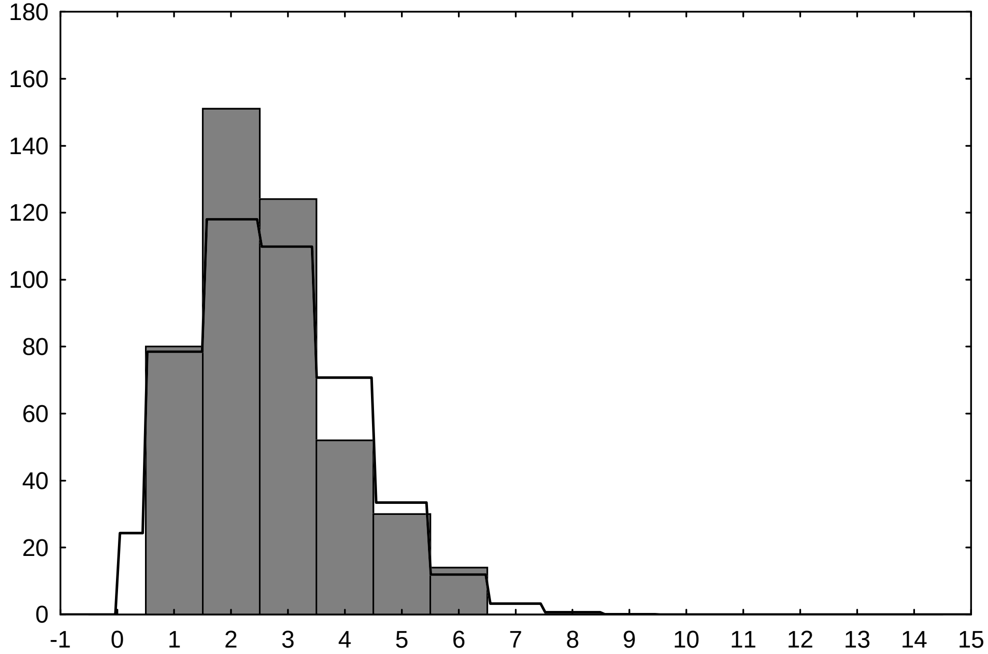


Moving-average chart ($w = 100$) on May, 17$^{\text{th}}$:



Out-of-control state is obvious, current value of p is significantly below p_0.

Histogram of log-in counts on May, 17th:



Probability p estimated as 0.17679, much below in-control value $p_0 = 0.38$.

The **explanation** for the unusual using behaviour observed is simple:

May 17th, 2005, was the tuesday after Whitsun. Whit Sunday and Whit Monday are holidays in whole Germany. On the tuesday after Whitsun, there are traditionally no lectures at the University of Würzburg.

Thank You
for Your Interest!

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics