# Statistische Kontrolle von Zähldatenprozessen mit Überdispersion

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics

This talk is based on the articles

**Weiß, C.H. (2008)**. *Modelling time series of counts with overdispersion*. Accepted for publication in Statistical Methods and Applications.

**Weiß, C.H. (2009)**. *The INARCH(1) Model for Overdispersed Time Series of Counts*. Submitted.

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

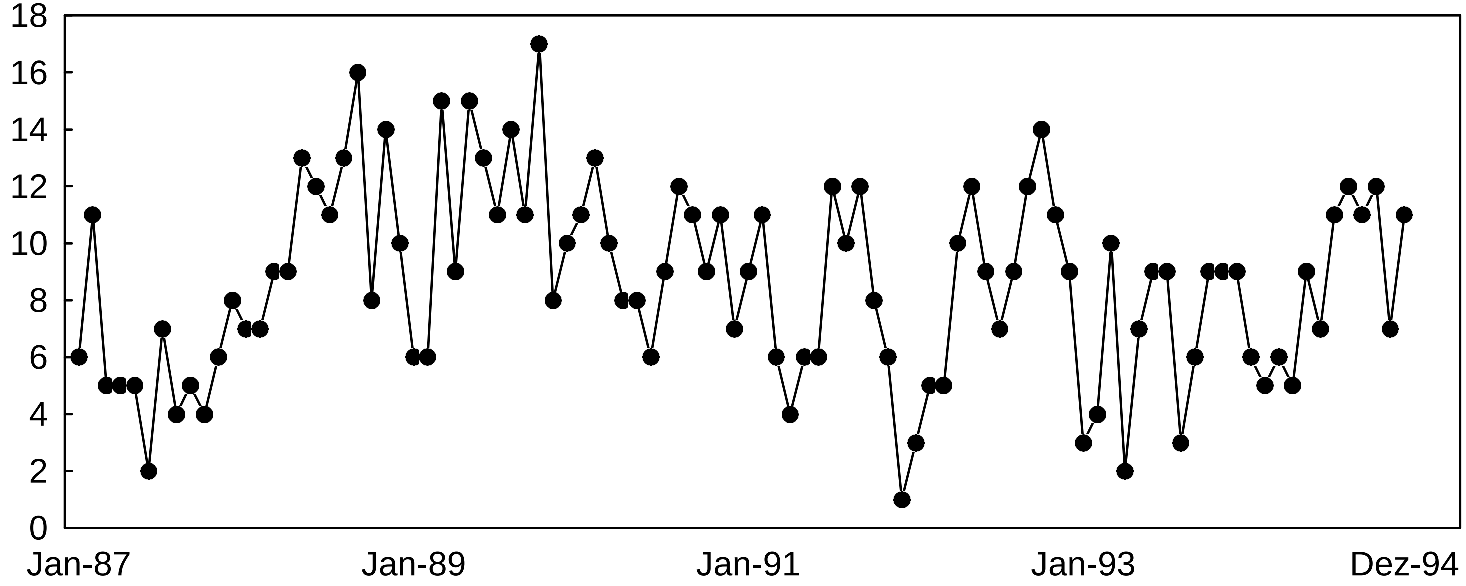
# Processes of Counts with Overdispersion

## Motivation

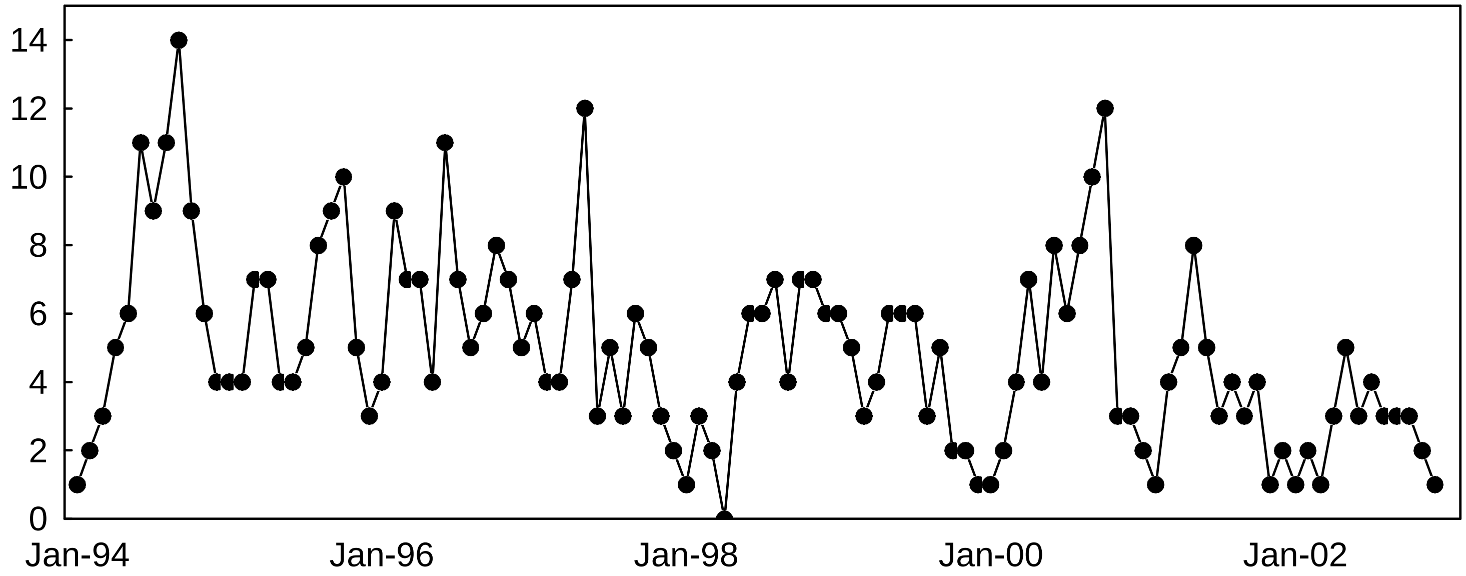
Processes of counts commonly observed in real-world applications. Examples from diverse fields in practice:

- insurance (e. g., time series of claim counts),
- economics (e. g., counts of price changes),
- statistical process control (e. g., counts of defects),
- traffic (e. g., counts of accidents),
- network monitoring (e. g., intrusion detection system),
- epidemiology (e. g., counts of diseases), and others.

**Example 1:** Monthly claims counts (1987 to 1994):
burn related injuries in heavy manufacturing industry.
Source: Freeland (1998).

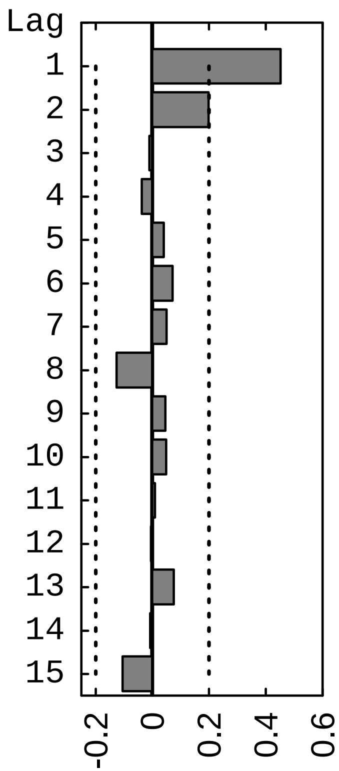


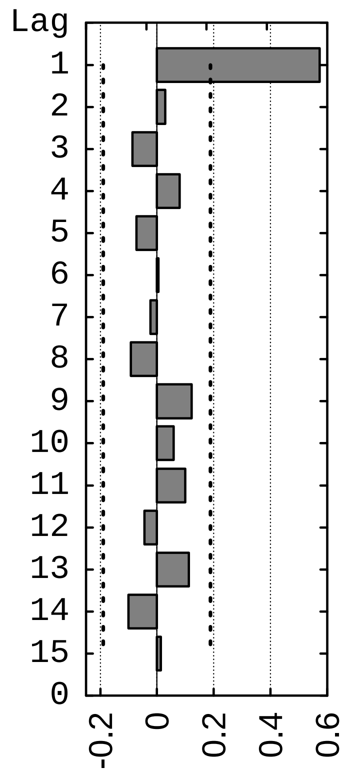
**Example 2:** Monthly strike data (1994 to 2002): number of work stoppages leading to 1000 workers or more being idle in effect in the period.
Source: U.S. Bureau of Labor Statistics.



## Analysis of both time series:

Partial autocorrelation function of

Example 1: 


Example 2: 

**Analysis of both time series:** (continued)

For both examples, AR(1)-like dependence structure

$\Rightarrow$ Popular Poisson INAR(1) model appropriate?

**Analysis of both time series:** (continued)

Marginal properties:

- Example 1: mean 8.60 and variance 11.36;
- Example 2: mean 4.94 and variance 7.92.

$\Rightarrow$ **Overdispersion** for both examples!

$\Rightarrow$ The popular Poisson INAR(1) model cannot be used!

Overdispersion commonly observed in practice.

Typical reasons:

- presence of positive correlation between monitored events (Friedman, 1993; Poortema, 1999; Paroli et al., 2000);
- variation in probability of monitored events (Heimann, 1996; Poortema, 1999; Christensen et al., 2003);
- further potential causes of overdispersion discussed by Jackson (1972).

Modeling time series of overdispersed counts:
**INGARCH models**, the *in*teger-valued *g*eneralized *a*uto-*r*egressive *c*onditional *h*eteroskedasticity models.

INGARCH models introduced by Heinen (2003),
further investigated by Ferland et al. (2006); Weiß (2008).

Defined by an ARMA-like recursion,
strictly stationary solution with finite first and second order moments exists (Ferland et al., 2006),
ARMA-like autocorrelation structure (Weiß, 2008).

# The INARCH(1) Model

## Definition & Properties

**Definition:**

Let $(X_t)_{\mathbb{Z}}$ be a process with range $\mathbb{N}_0 = \{0, 1, \ldots\}$,

let $\beta > 0$ and $0 < \alpha < 1$.

$(X_t)_{\mathbb{Z}}$ is said to follow an **INARCH(1) model**

if $X_t$, conditioned on $X_{t-1}, X_{t-2}, \ldots$,

is Poisson distributed according to $Po(\beta + \alpha \cdot X_{t-1})$.

**Basic properties:**

- Stationary Markov chain with transition probabilities

$$\begin{aligned} p_{i|j} &:= P(X_t = i \mid X_{t-1} = j) \\ &= \exp\left(-\beta - \alpha \cdot j\right) \cdot \tfrac{(\beta+\alpha\cdot j)^i}{i!} \ > 0; \end{aligned}$$

- autocorrelation function $\rho_X(n) := Corr[X_t, X_{t-n}] = \alpha^n$.

**Proposition: (Marginal Cumulants)**

The cumulants follow recursively from

$$\kappa_1 = \tfrac{\beta}{1-\alpha}, \qquad \kappa_n = -(1-\alpha^n)^{-1} \cdot \textstyle\sum_{j=1}^{n-1} s_{n,j} \cdot \kappa_j \quad \text{for } n \geq 2,$$

where $s_{n,j}$ are Stirling numbers of first kind:

$$s_{n,0} = 0 \quad \text{and} \qquad s_{n,n} = 1 \quad \text{for } n \geq 1,$$
$$s_{n+1,j} = s_{n,j-1} - n \cdot s_{n,j} \qquad \text{for } j = 1, \ldots, n \text{ and } n \geq 1.$$

**Proposition: (Marginal Cumulants)** (continued)

In particular,

$$\kappa_1 = \tfrac{\beta}{1-\alpha} = E[X_t], \quad \kappa_2 = \tfrac{\beta}{(1-\alpha)(1-\alpha^2)} = V[X_t],$$

i. e., **overdispersion**,

$$\kappa_3 = \frac{1+2\alpha^2}{1-\alpha^3}\cdot\kappa_2, \quad \kappa_4 = \frac{1+6\alpha^2+5\alpha^3+6\alpha^5}{(1-\alpha^3)(1-\alpha^4)}\cdot\kappa_2,$$

i. e., skewness and excess of $X_t$ are given by
$\frac{1+2\alpha^2}{1+\alpha+\alpha^2}\cdot\sqrt{\frac{1+\alpha}{\beta}}$ and $\frac{1+6\alpha^2+5\alpha^3+6\alpha^5}{\beta(1+\alpha+\alpha^2)(1+\alpha^2)}$, respectively.

## Estimation of Parameters:

- conditional maximum likelihood approach:

$$p_{i|j} \;=\; \exp\left(-\beta-\alpha\cdot j\right)\cdot\frac{(\beta+\alpha\cdot j)^i}{i!};$$

- conditional least squares approach:

$$E[X_t \mid X_{t-1}=x_{t-1}] \;=\; \beta+\alpha\cdot x_{t-1};$$

- method of moments:

$$\mu_X \;=\; \frac{\beta}{1-\alpha}, \qquad \rho_X(1) \;=\; \alpha.$$

INARCH(1) model performs very well for above examples:

- **Example 1:**
  ML-estimates $\hat{\beta} = 4.3796$ and $\hat{\alpha} = 0.4911$,
  model mean 8.61 and variance 11.34,
  empirical mean 8.60 and variance 11.36.

- **Example 2:**
  ML-estimates $\hat{\beta} = 1.8114$ and $\hat{\alpha} = 0.6364$,
  model mean 4.98 and variance 8.37,
  empirical mean 4.94 and variance 7.92.

# Monitoring INARCH(1) Processes

Approaches

INARCH(1) process is AR(1)-like process of counts
$\Rightarrow$ try to adapt approaches developed for Poisson INAR(1) processes (Weiß, 2007, 2009; Weiß & Testik, 2009).

**Most basic approach:**
$c$ chart with lower and upper control limits LCL and UCL, which monitors the observed counts $X_t$ directly.

Since INARCH(1) process is Markov chain, ARLs can be computed with Markov chain approach of Brook & Evans (1972).

**Markov chain approach of Brook & Evans (1972):**

Let in-control model $(\beta_0, \alpha_0)$ and $LCL, UCL \in \mathbb{N}_0$ be fixed.

$\Rightarrow$ in-control states $LCL, \ldots, UCL$.

Corresponding true transition probabilities

$$p_{i|j} \;=\; \exp\left(-\beta - \alpha \cdot j\right) \cdot \frac{(\beta+\alpha\cdot j)^i}{i!}$$

summarized in matrix

$$\mathbf{Q} \;:=\; \begin{pmatrix} p_{LCL|LCL} & \cdots & p_{UCL|LCL} \\ \vdots & & \vdots \\ p_{LCL|UCL} & \cdots & p_{UCL|UCL} \end{pmatrix}.$$

**MC approach:** (continued)

The components $\mu_i$ of solution of $(\mathbf{I}-\mathbf{Q})\boldsymbol{\mu} = \mathbf{1}$ are conditional ARLs, conditioned on event that process started in state $i$.

$\Rightarrow$ Overall ARL given by

$$\text{ARL} \;=\; 1 + \textstyle\sum_{i=LCL}^{UCL} \mu_i \cdot P(X_t = i).$$

**Problem:**

Explicit expression for marginal probabilities $p_i := P(X_t = i)$ of INARCH(1) process not known!

**First solution:**

$(X_t)_{\mathbb{Z}}$ is ergodic Markov chain, it follows that

$$p_i \;=\; \lim_{n\to\infty}\; p_{i|j}(n) \qquad \text{for all } i,j\in\mathbb{N}_0,$$

where $n$-step transition probabilities

$$p_{i|j}(n) := P(X_t = i \mid X_{t-n} = j)$$

follow recursively via

$$p_{i|j}(n) \;=\; \sum_{r=0}^{\infty}\; p_{i|r}\cdot p_{r|j}(n-1).$$

**First solution:** (continued)

These relations allow to determine marginal probabilities numerically:

Choosing $M, N \in \mathbb{N}$ sufficiently large, we approximate

$$p_i \approx p_{i|j}(N), \qquad \text{where } p_{i|j}(n) \approx \sum_{r=0}^{M} p_{i|r} \cdot p_{r|j}(n-1)$$

for arbitrary $i, j \in \mathbb{N}_0$, e. g., choose $j := \lceil \mu_X \rceil$.

**Disadvantage:** computationally rather intensive, requires appropriate choice of $M, N$.

# Poisson-Charlier Expansion

Background

Probability generating function (pgf) of $X$: $p_X(z) := E[z^X]$.

**Factorial cumulant generating function (fcgf):**

$$k_X(z) \ := \ \ln\big(p_X(1+z)\big) \ = \ \ln E[(1+z)^X] \ =: \ \sum_{r=1}^{\infty} \frac{\kappa_{(r)}}{r!}\cdot z^r,$$

where coefficients $\kappa_{(r)}$ referred to as **factorial cumulants**.

Factorial cumulants related to usual cumulants via

$$\kappa_{(n)} \ = \ \textstyle\sum_{j=1}^{n} \ s_{n,j}\cdot\kappa_j. \qquad (s_{n,j}\text{: Stirling numbers})$$

**Example:** Poisson distribution $Po(\lambda)$, then $k_X(z) = \lambda z$, i. e., $\kappa_{(1)} = \kappa_1 = \lambda$ and $\kappa_{(r)} = 0$ for $r \geq 2$.

If fcgf of $X$ known, then pgf

$$p_X(z) \;=\; \exp\big(k_X(z-1)\big) \;=\; \exp\big(\textstyle\sum_{r=1}^{\infty}\,\frac{\kappa_{(r)}}{r!}\cdot(z-1)^r\big).$$

**Idea:** Approximate true pgf of $X$ by $m^{\text{th}}$ order approximation

$$p_X(z) \;\approx\; \exp\big(\textstyle\sum_{r=1}^{m}\,\frac{\kappa_{(r)}}{r!}\cdot(z-1)^r\big).$$

If $X$ Poisson distributed, then first order approximation already gives exact pgf.

**Poisson-Charlier expansion** of Barbour (1987) further refinement of this approach.

Let $\pi_i := e^{-\kappa_1} \cdot \kappa_1^i / i!$ denote Poisson probabilities.
Let $\nabla$ denote difference operator: $\nabla \pi_i = \pi_i - \pi_{i-1}$.
Then $m^{\text{th}}$ order **Poisson-Charlier approximation** of true probability $p_i$ given by $f_m(\nabla) \cdot \pi_i$, where $f_m$ is $(m-1)^{\text{th}}$ order Taylor polynomial around $z = 0$ and evaluated in $z = 1$ of

$$f(z, \nabla) \; := \; \exp\left(\frac{1}{z} \cdot \sum_{r=2}^{\infty} \frac{\kappa_{(r)}}{r!} \cdot (-z\nabla)^r\right).$$

The first four Poisson-Charlier approximations:

$$\begin{aligned} f_1(\nabla) &= 1, \\ f_2(\nabla) &= 1 + \tfrac{1}{2}\kappa_{(2)}\nabla^2, \\ f_3(\nabla) &= 1 + \tfrac{1}{2}\kappa_{(2)}\nabla^2 - \tfrac{1}{6}\kappa_{(3)}\nabla^3 + \tfrac{1}{8}\kappa_{(2)}^2\nabla^4, \\ f_4(\nabla) &= 1 + \tfrac{1}{2}\kappa_{(2)}\nabla^2 - \tfrac{1}{6}\kappa_{(3)}\nabla^3 + \big(\tfrac{\kappa_{(2)}^2}{8} + \tfrac{\kappa_{(4)}}{24}\big)\nabla^4 \\ &\quad - \tfrac{1}{12}\kappa_{(2)}\kappa_{(3)}\nabla^5 + \tfrac{1}{48}\kappa_{(2)}^3\nabla^6. \end{aligned}$$

So only knowledge about first few factorial cumulants of $X$ required!

**Proposition: (Marginal Factorial Cumulants)**

Factorial cumulants of INARCH(1) process determined from usual cumulants via

$$\kappa_{(1)} = \kappa_1, \qquad \kappa_{(n)} = \alpha^n \cdot \kappa_n \quad \text{for } n \geq 2.$$

# Poisson–Charlier Expansion

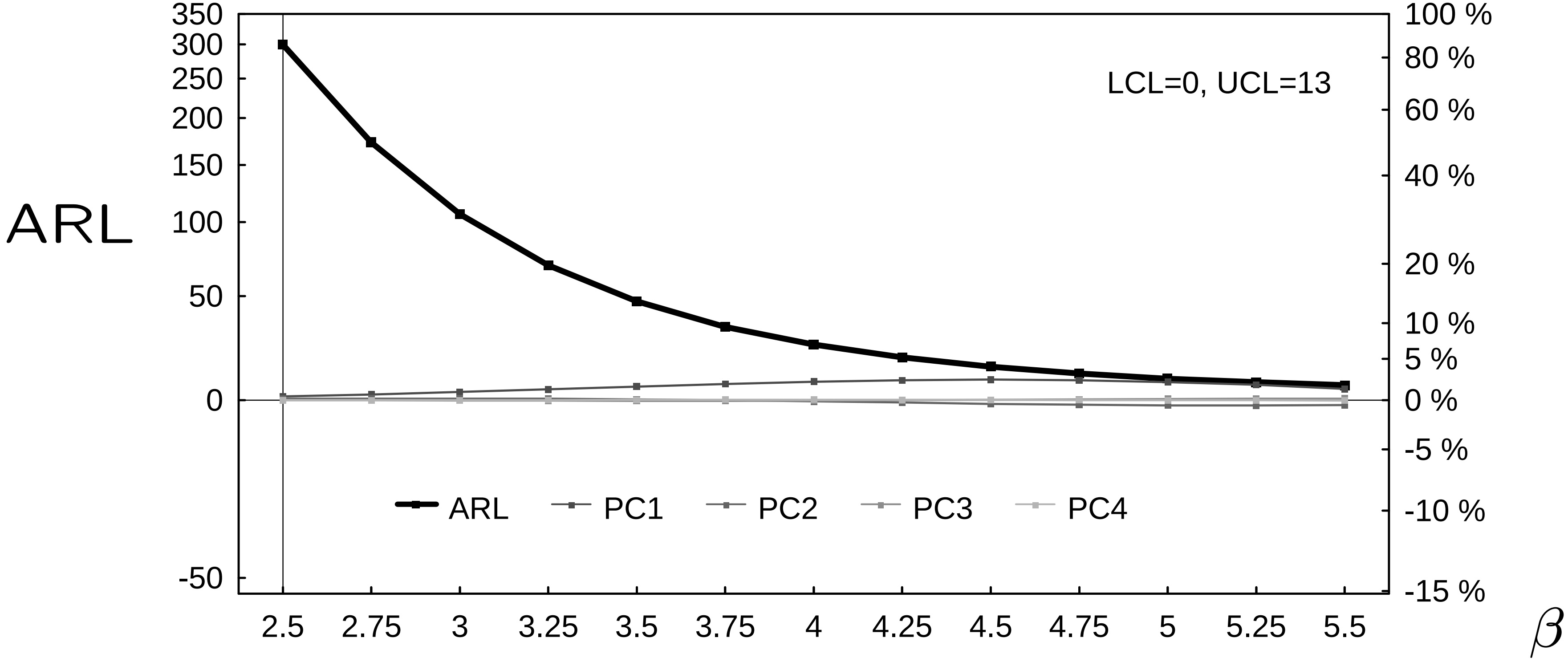
Performance

We investigate performance of Poisson-Charlier approximations by considering effect on ARL of $c$ chart.

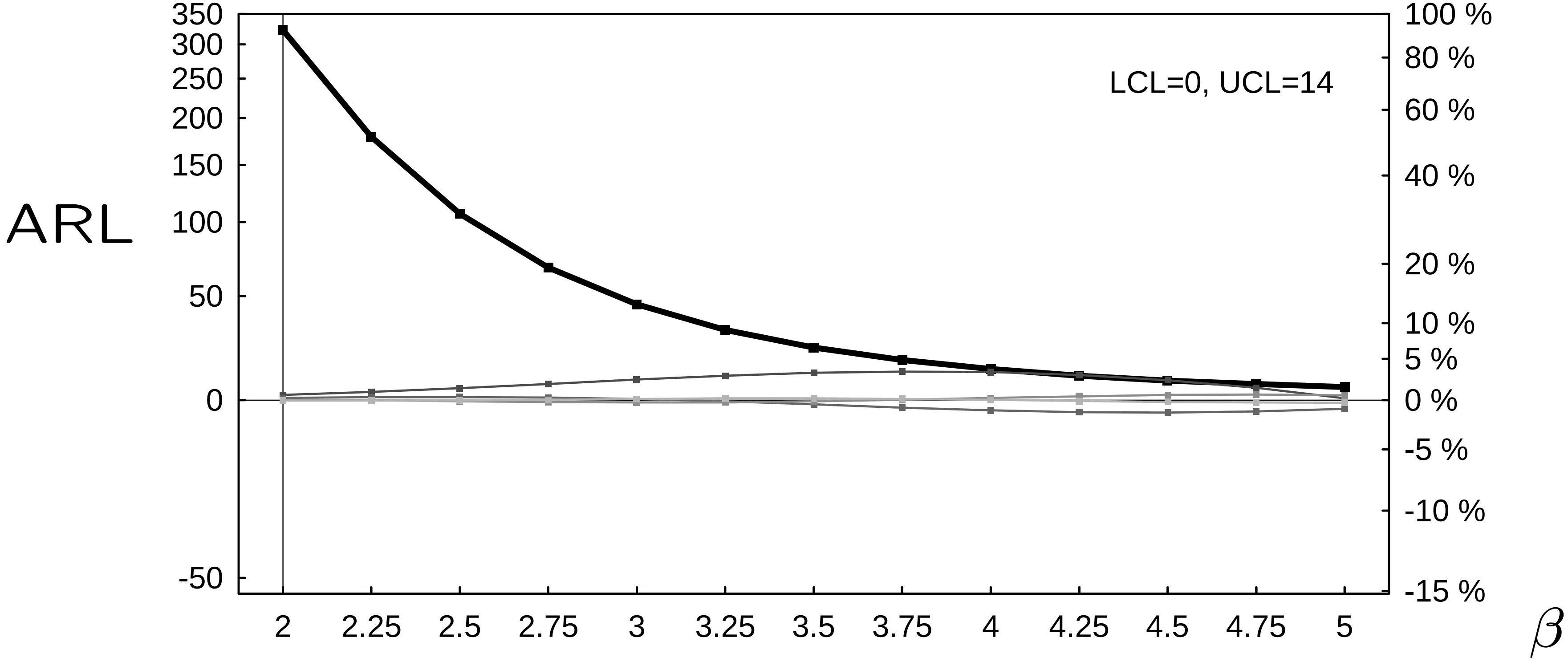
We consider approximations up to order 4, since higher order approximations become too complex for practice.

Some illustrative graphs in the following, where shifts in $\beta$ compared to $\beta_0$ are considered, while $\alpha = \alpha_0$.

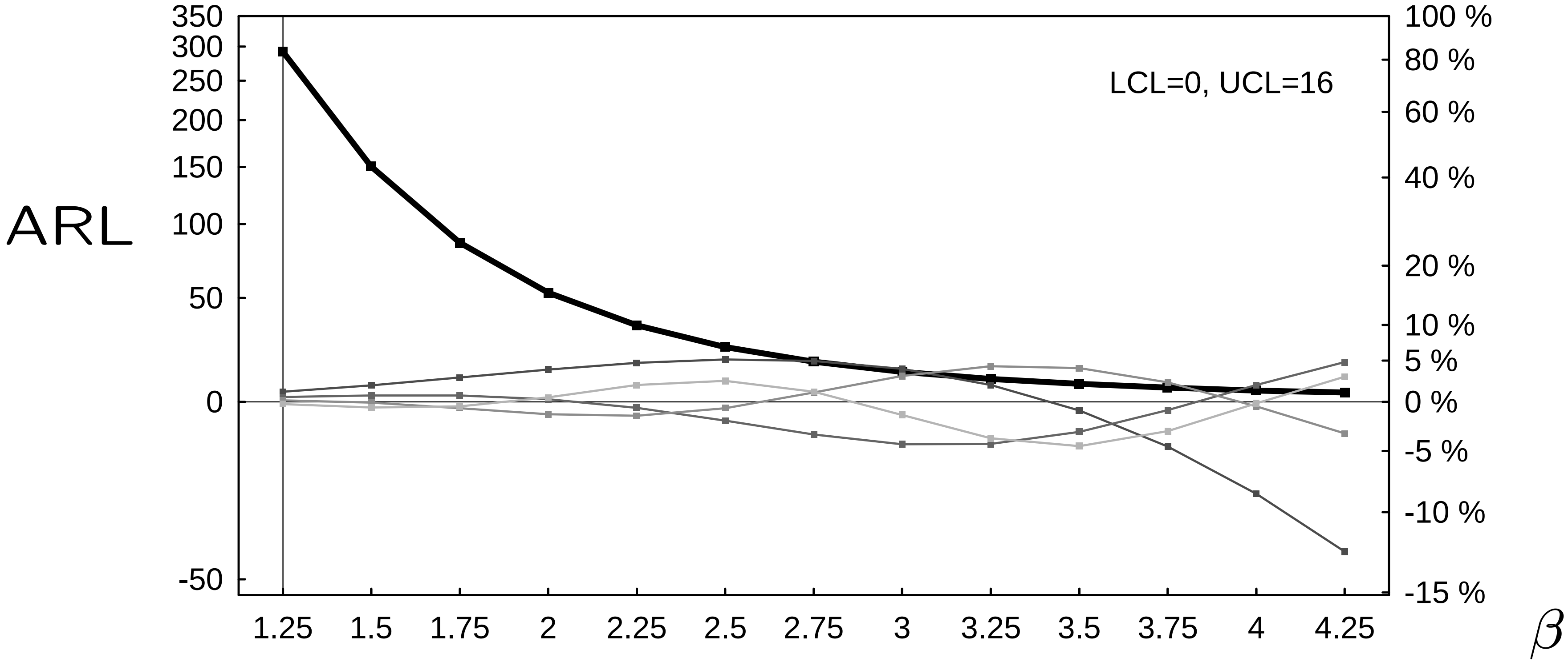
ARL($\beta$) of $c$ chart and relative errors of Poisson-Charlier approximations (PC$n$) for $(\beta_0, \alpha_0) = (2.5, 0.5)$:



ARL($\beta$) of $c$ chart and relative errors of Poisson-Charlier approximations (PC$n$) for $(\beta_0, \alpha_0) = (2, 0.6)$:



ARL($\beta$) of $c$ chart and relative errors of Poisson-Charlier approximations (PC$n$) for $(\beta_0, \alpha_0) = (1.25, 0.75)$:



It becomes clear that for $\alpha_0 \leq 0.5$, any Poisson-Charlier approximation of order $\geq 2$ leads to a satisfactory approximation of the ARLs.

For $\alpha_0 = 0.6$, at least the approximations of order $\geq 3$ can be used, while these approximations lead to errors between $-5\,\%$ and $+5\,\%$ for $\alpha_0 = 0.75$.

- INARCH(1) model: simple and parsimoniously parametrized model for time series of overdispersed counts.

- Explicit expressions for marginal (factorial) cumulants, autocorrelation function, transition probabilities, but not for marginal probabilities.

- Approximate ARLs of $c$ chart through approximation of marginal distribution via Poisson-Charlier expansion. PC approximation better than Poisson approximation, really satisfactorily only for moderate autocorrelation.

# Thank You for Your Interest!

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics