# Statistical Models for Defective Count Data

Gerhard Neubauer[a], Gordana Đuraš[a], and Herwig Friedl[b]

[a]Statistical Applications, Joanneum Research, Graz, Austria
[b]Institute of Statistics, University of Technology, Graz, Austria

Statistik Tage 2011, Graz, September 7th-9th

# Introduction

# Defective Count Data

Defective: too much and/or too few is counted, recorded, reported ...

Most prominent example: Criminology

- Crimes associated with shame (sexual offences, domestic violence)
- Theft of low-value goods

Likely to be not reported to the police, official numbers are too low

$\rightarrow$ **Under-Reporting**

# Defective Count Data

Less considered:
Insurance companies suspect that theft may be reported, but only to make an insurance claim

Most popular with bicycles, skiing equipment

Likely to be reported to the police, official numbers are too high

$\longrightarrow$ **Over-Reporting**

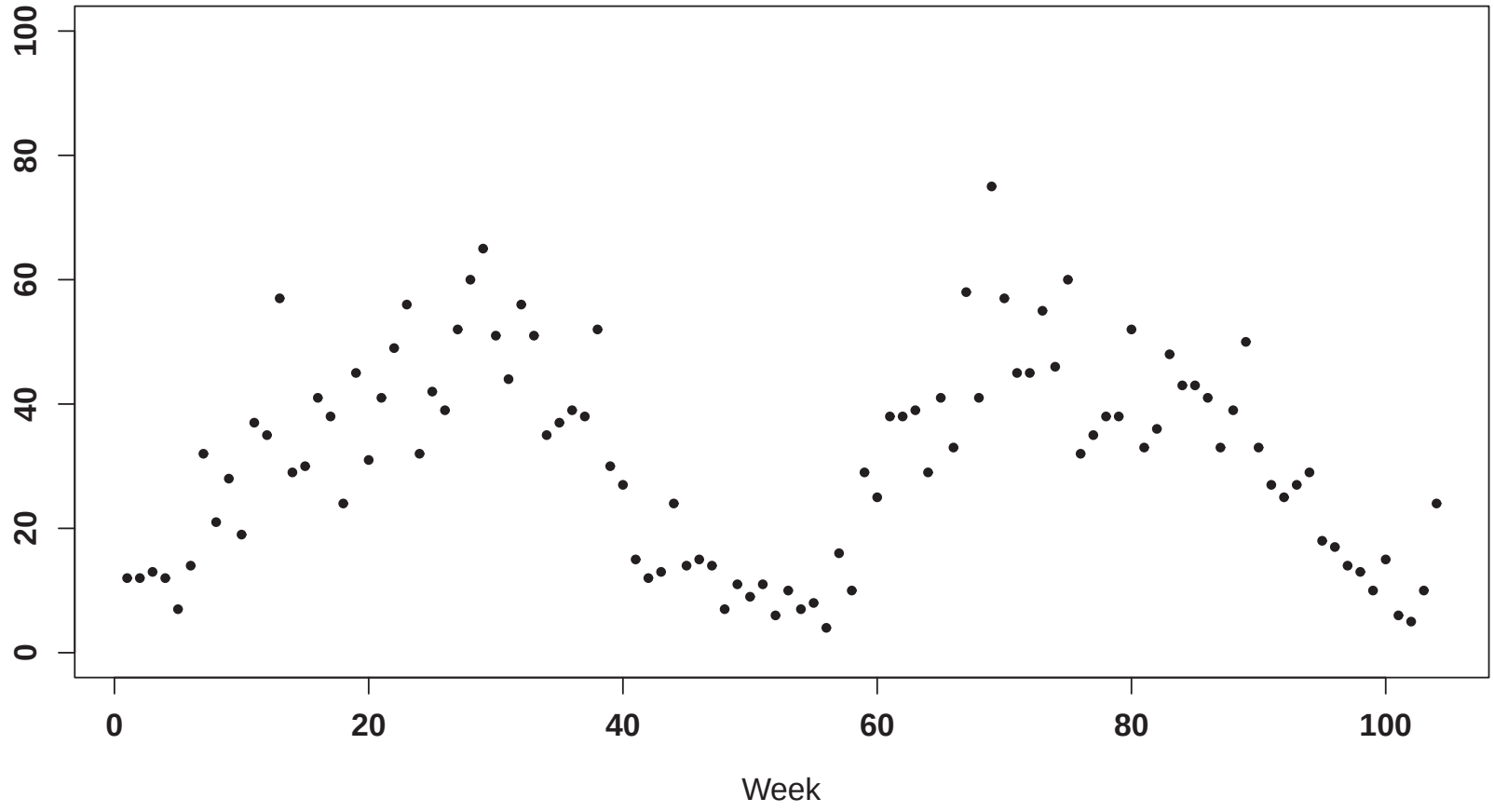
More precisely:

$\longrightarrow$ **Under-Reporting fraud**

$\longrightarrow$ **Over-Reporting theft**

# Criminology: Bicycle Theft Data

Weekly counts of bicycle theft in an Austrian region



# Bicycle Theft ?

# Defective Count Data

Health data

Registers for infectious diseases (HIV), chronic diseases (diabetes)

Cause of death registers (heart attack)

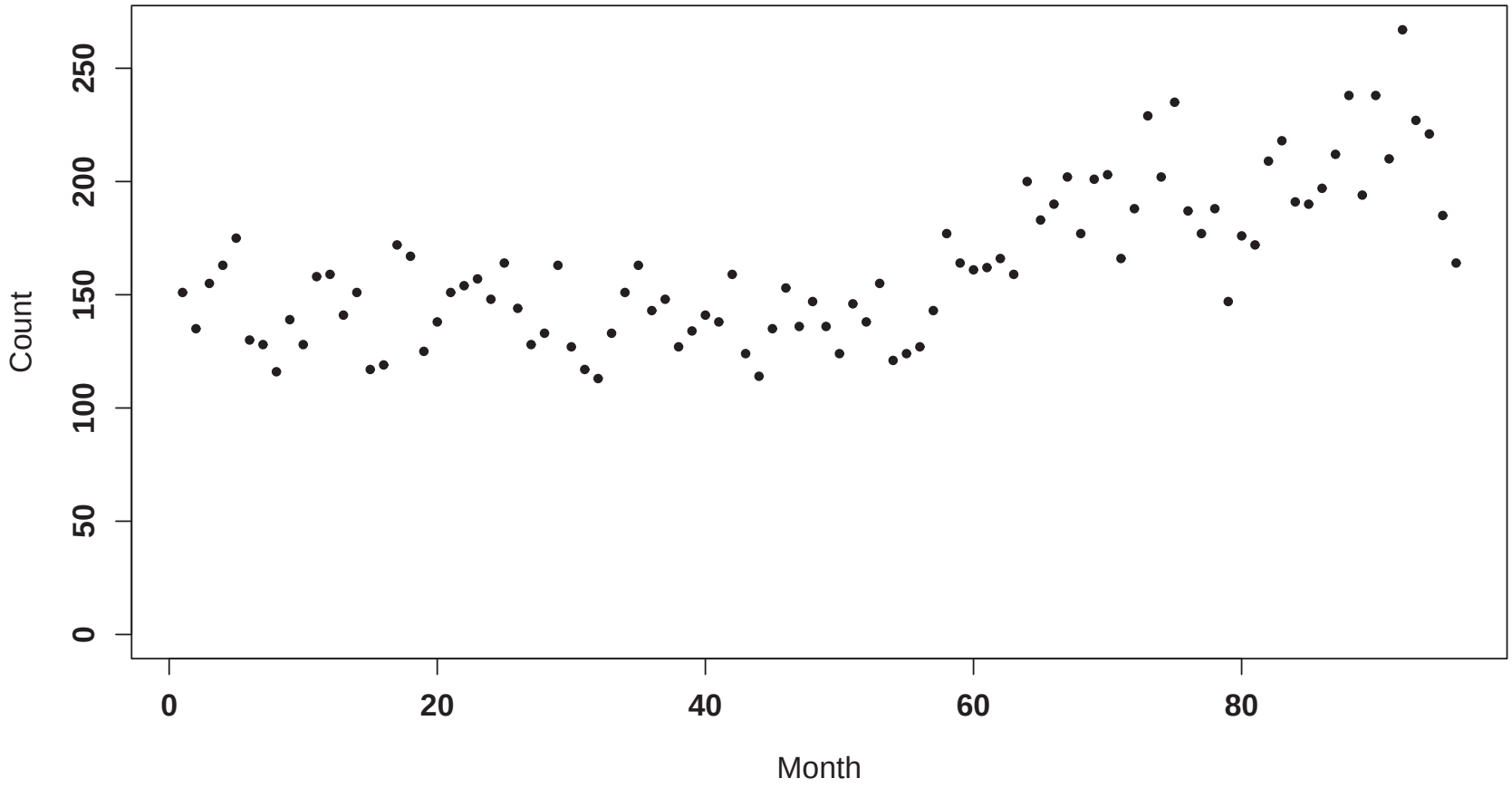
Any miss-classification will cause both errors

$\longrightarrow$ **Under-Reporting in the right category**

$\longrightarrow$ **Over-Reporting in the wrong category**

# Health: Heart Attack Data

Monthly counts of heart attack discharges from Styrian hospitals



# Defective Count Data

- Criminology
- Health data
- Insurance data: traffic accidents with minor damage
- Production: number of defective goods in production
- Warranty: number of goods sent back for warranty claim
- ...

Any sample of count data may be defective due to recording failures

# Under-Reporting

**How to estimate the total number of cases?**

# Bernoulli Sampling

One observation

Case reported

| Yes | No |
|---|---|
| $R$ | $1 - R$ |

$\lambda$ observations

Case reported

| Yes | No |
|---|---|
| $Y$ | $D$ |

$R \sim \text{Bernoulli}(\pi)$

$$Y = \sum_{i=1}^{\lambda} R_i \sim \text{Binomial}(\lambda, \pi)$$

$Y$ . . . (random) number of reported cases

$D$ . . . (random) number of unreported cases

$\lambda$ . . . total number of cases

$\pi$ . . . reporting probability

# Binomial Model

Case reported

| Yes | No |
|---|---|
| $Y$ | $D$ |

$$Y = \sum_{i=1}^{\lambda} R_i \sim \text{Binomial}(\lambda, \pi)$$

$$\begin{aligned} \mathrm{E}(Y) &= \mu = \lambda\pi \\ \mathrm{var}(Y) &= \sigma^2 = \lambda\pi(1-\pi) \end{aligned}$$

**Both $\lambda$ and $\pi$ have to be estimated**

# Binomial Model

Case reported

| Yes | No |
|---|---|
| $Y$ | $D$ |

$$Y = \sum_{i=1}^{\lambda} R_i \sim \text{Binomial}(\lambda, \pi)$$

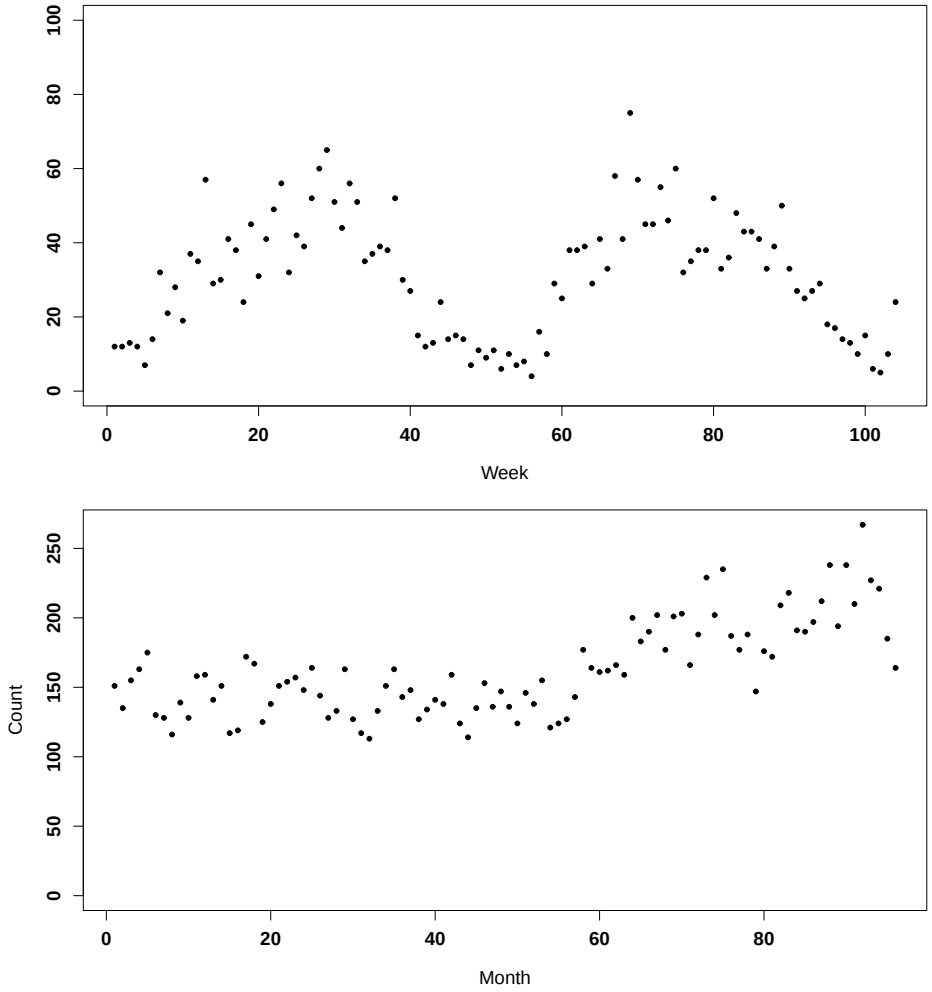
$$\begin{aligned} \mathrm{E}(Y) &= \mu = \lambda\pi \\ \mathrm{var}(Y) &= \sigma^2 = \lambda\pi(1-\pi) \end{aligned}$$

**Both $\lambda$ and $\pi$ have to be estimated**

**No longer a member of 1-parameter exponential family**

**Limitation to data with $s^2 < \overline{y}$**

# Is iid Assumption Appropriate?




**Trend/Seasonality** $\Rightarrow$ **Regression approach**

# Regression Approach

Consider $Y_t \overset{ind}{\sim} \text{Binomial}(\lambda_t, \pi), \quad t = 1, \ldots, T,$
where

$$\begin{aligned} \lambda_t &= \exp(\mathbf{x}_t'\boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \mathbb{R}^d, \\ \pi &= \frac{\exp(\alpha)}{1+\exp(\alpha)}, \quad \alpha \in \mathbb{R}, \end{aligned}$$

to ensure $\lambda_t > 0$ and $0 < \pi < 1$

Likelihood contribution of $y_t$

$$L(\alpha, \boldsymbol{\beta}|y_t) = \binom{\lambda_t}{y_t} \pi^{y_t}(1-\pi)^{\lambda_t - y_t}$$

# Maximum Likelihood Estimation

Sample Log-Likelihood

$$\begin{aligned}\ell(\alpha, \boldsymbol{\beta}|\mathbf{y}) &= \sum_{t=1}^{T} \log L(\alpha, \boldsymbol{\beta}|y_t) \\ &= \sum_{t=1}^{T} \log \left[ \binom{\lambda_t}{y_t} \pi^{y_t} (1-\pi)^{\lambda_t - y_t} \right]\end{aligned}$$

Score and observed Information are analytically evaluated

Newton-Raphson procedure till convergence

# Maximum Likelihood Estimation

Sample Log-Likelihood

$$\begin{aligned}\ell(\alpha, \boldsymbol{\beta}|\mathbf{y}) &= \sum_{t=1}^{T} \log L(\alpha, \boldsymbol{\beta}|y_t) \\ &= \sum_{t=1}^{T} \log \left[ \binom{\lambda_t}{y_t} \pi^{y_t} (1-\pi)^{\lambda_t - y_t} \right]\end{aligned}$$

Score and observed Information are analytically evaluated

Newton-Raphson procedure till convergence

**Still problems with overdispersion!**

# Randomized Parameter Models

# Beta-Binomial Model

## Randomize reporting probability $\pi$

$$\begin{aligned} Y_t|P_t &\sim \text{Binomial}(\lambda_t, P_t) \\ P_t &\sim \text{Beta}(\gamma, \delta) \\ Y_t &\sim \text{Beta-Binomial}(\lambda_t, \gamma, \delta) \end{aligned}$$

We use the parameterization

$$\theta = \gamma + \delta \qquad \text{and} \qquad \pi = \frac{\gamma}{\gamma + \delta} = E(P_t)$$

Hence $\mathrm{E}(Y_t) = \mu_t = \lambda_t \pi$ and $\mathrm{var}(Y_t) = \mu_t(1-\pi)\phi$,

where $\phi = \frac{\lambda+\gamma+\delta}{1+\gamma+\delta} \geq 1$

## Randomize total number of cases $\lambda$

$$\begin{aligned} Y_t|L_t &\sim \text{Binomial}(L_t, \pi) \\ L_t &\sim \text{Poisson}(\lambda_t) \\ Y_t &\sim \text{Poisson}(\lambda_t \pi) \end{aligned}$$

Model not identified

Winkelmann (2000): Pogit Model

$$Y_t \sim \text{Poisson}(\lambda_t \pi_t)$$

$$\lambda_t = \exp(\mathbf{x}_t' \boldsymbol{\beta}) \text{ and } \pi_t = \frac{\exp(\mathbf{z}_t' \boldsymbol{\alpha})}{1 + \exp(\mathbf{z}_t' \boldsymbol{\alpha})}$$

with two (disjoint) sets of regressors $\mathbf{x}_t$ and $\mathbf{z}_t$

# Negative Binomial Model

## Randomize total number of cases $\lambda$

$$\begin{aligned} Y_t|L_t &\sim \text{Binomial}(L_t, \pi) \\ L_t|K_t &\sim \text{Poisson}(K_t\lambda_t) \\ K_t &\sim \text{Gamma}(\omega_t, \omega_t) \\ Y_t &\sim \text{Negative Binomial}(\omega_t, 1-\pi) \end{aligned}$$

where $\omega_t = \lambda_t(1-\pi)$ is the expected number of unreported cases

$$\mathrm{E}(Y_t) = \mu_t = \lambda_t\pi \qquad\qquad \mathrm{var}(Y_t) = \mu_t + \frac{\mu_t^2}{\omega_t}$$

# Beta-Poisson Model

## Randomize $\pi$ and $\lambda$

$$\begin{aligned} Y_t|L_t, P_t &\sim \text{Binomial}(L_t, P_t) \\ L_t &\sim \text{Poisson}(\lambda_t) \\ P_t &\sim \text{Beta}(\gamma, \delta) \\ Y_t &\sim \text{Beta-Poisson}(\lambda_t, \gamma, \delta) \end{aligned}$$

We use the parameterization

$$\theta = \gamma + \delta \qquad \text{and} \qquad \pi = \frac{\gamma}{\gamma + \delta} = E(P_t)$$

Hence $\mathrm{E}(Y_t) = \mu_t = \lambda_t \pi$ and $\mathrm{var}(Y_t) = \mu_t \phi_t$,

where $\phi_t = 1 + \frac{\lambda_t(1-\pi)}{1+\gamma+\delta} \geq 1$

# Estimation and Inference

# Estimation

We use **Maximum Likelihood (ML)**, where the solution of score equations

$$\frac{\partial}{\partial \alpha} \log \prod_t L\left(\Omega_M | y_t, \mathbf{x}_t\right) = 0 \text{ and } \frac{\partial}{\partial \beta} \log \prod_t L\left(\Omega_M | y_t, \mathbf{x}_t\right) = 0$$

$\Omega_M$ the parameter vector of given model

is found by the **Newton-Raphson algorithm**

For the Beta-mixtures we also use the Hybrid algorithm

Cycle between

1. ML for $\alpha, \beta$ given $\theta$, and
2. MoM for $\theta$ given $\alpha, \beta$

## Choice of starting values sometimes crucial

# Inference

**Normality of parameter estimates**

Simulation studies: for reasonable settings $\hat{\alpha}$ and $\hat{\beta}$ approximately normal

Problems when

$\pi \to 0$, i.e. Poisson limit

$\pi \to 1$, perfect reporting system

**Inference within distribution** - usual model selection methods

e.g. $t$-values, LRT,...

# Inference

**Inference between distributions** - Non-nested testing (Allcroft and Glasbey, 2003)

Comparison of various models $\mathbf{M} = (M_k), \quad k = 1, \ldots, K$

1. find $\hat{\theta}_k$ for data $\mathbf{y} = (y_t)$ by optimizing GoF criterion $\mathbf{c} = (c_k(\mathbf{y}))$

2. simulate samples $\mathbf{y}^{(s)}(\hat{\theta}_k)$, $s = 1, \ldots, S$, (e.g. $S = 100$) and obtain all $S$ matrices $\mathbf{C}^{(s)}$ of dimension $K \times K$

3. obtain mean $\bar{\mathbf{C}}$ of $S$ matrices and compare $\mathbf{c}$ to the $k$-th column $\bar{\mathbf{c}}_k$ by multivariate normality assumption

   If $D_k \sim \chi^2_K$, $D_k = (\mathbf{c} - \bar{\mathbf{c}}_k)' V_k^{-1} (\mathbf{c} - \bar{\mathbf{c}}_k) \Rightarrow k$th model correct

# Real Data Applications

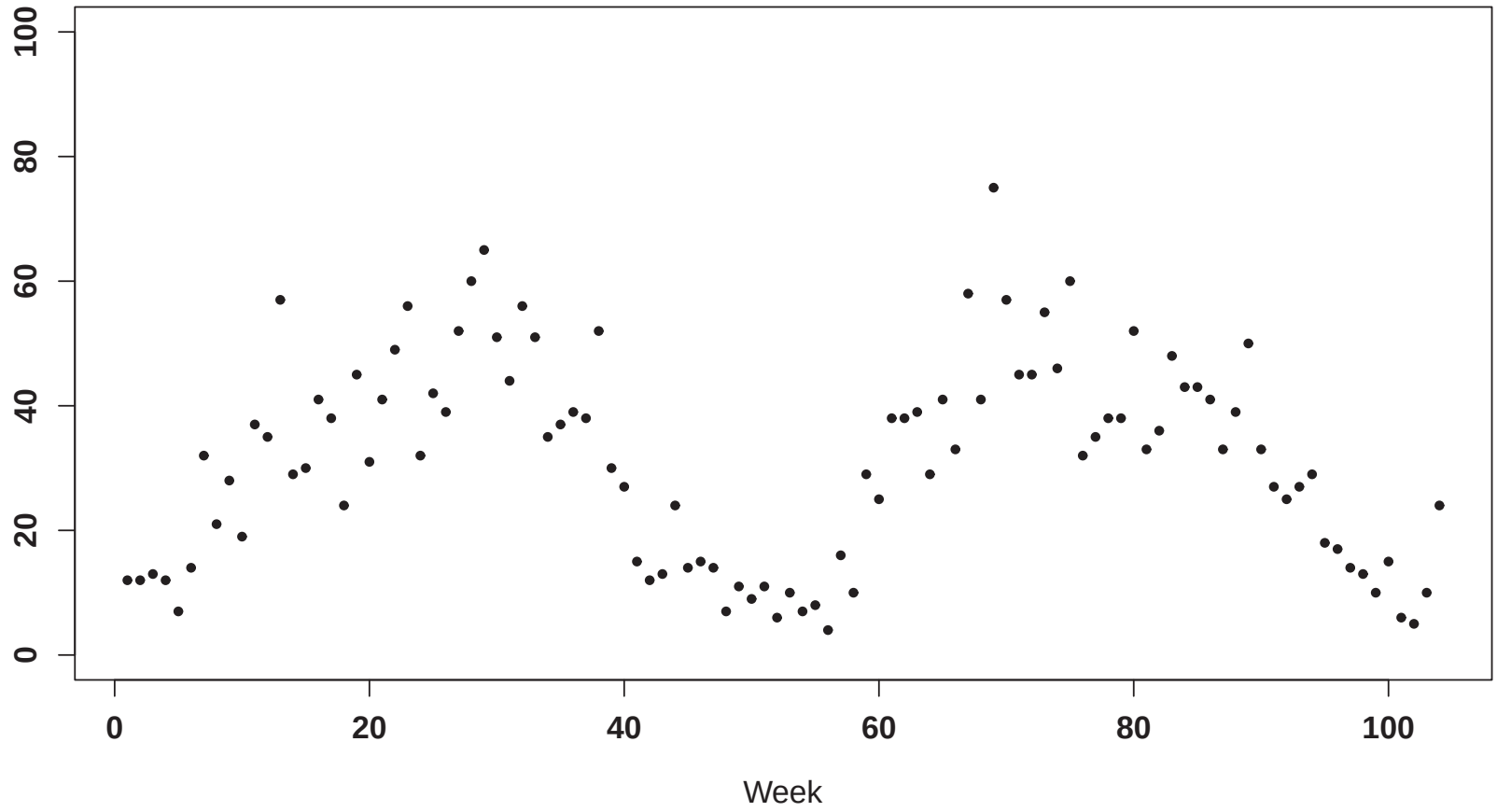
# Bicycle Theft

| Distribution | $\log L$ | Pearson | BIC | $D$ | $p(D)$ | $\hat{\pi}$ |
|---|---|---|---|---|---|---|
| Negative-Binomial | -728.91 | 206.49 | 1521.63 | 1.92 | 0.59 | 0.61 |
| Beta-Binomial | -732.87 | 204.79 | 1529.57 | 9.59 | 0.02 | 0.32 |
| Beta-Poisson | -735.39 | 197.97 | 1534.60 | 9.85 | 0.02 | 0.63 |

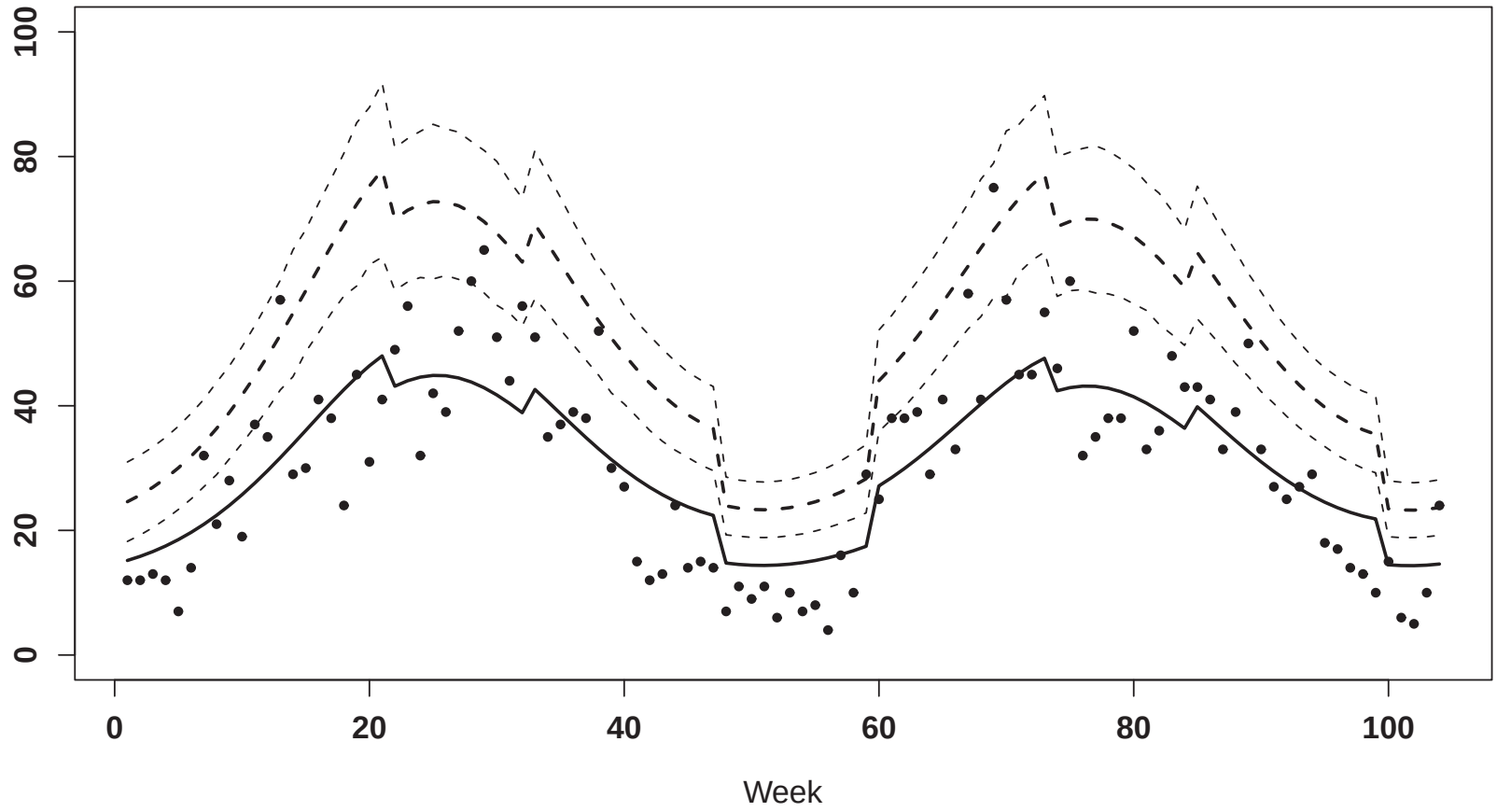
Model selected: Negative Binomial

$$\hat{\pi} = 0.61,\ p(D) = 0.59$$

# Bicycle Theft Data



# Bicycle Theft Model




Estimated mean (solid), estimated total number of thefts with confidence interval (dashed)

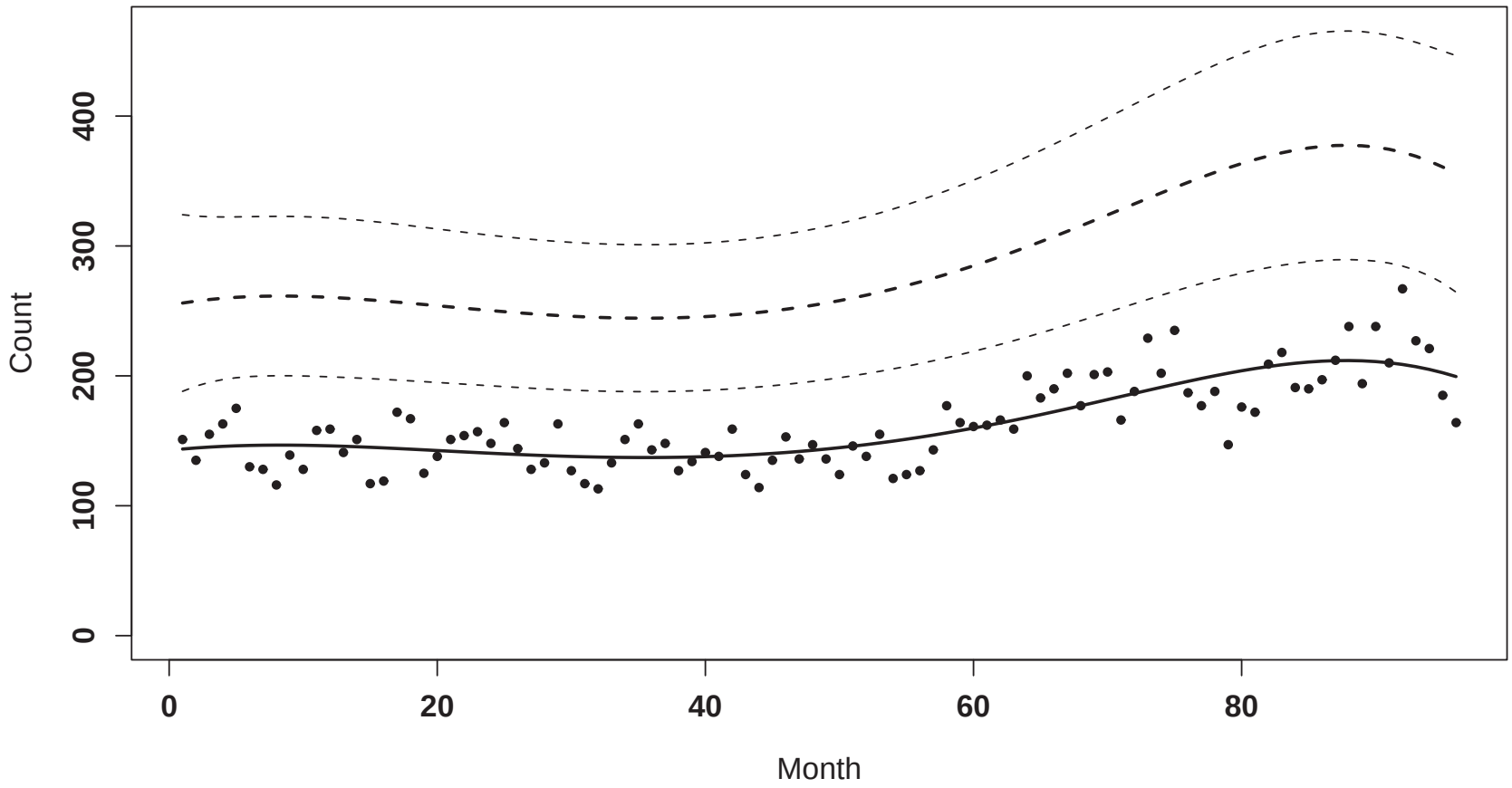
# Heart Attack Data

| Distribution | $\log L$ | Pearson | BIC | $D$ | $p(D)$ | $\hat{\pi}$ |
|---|---|---|---|---|---|---|
| Negative Binomial | -419.23 | 95.56 | 870.41 | 1.75 | 0.63 | 0.55 |
| Beta-Binomial | -417.64 | 95.60 | 871.79 | 77.61 | <e-03 | 0.62 |
| Beta-Poisson | -419.82 | 88.00 | 876.16 | 36.17 | <e-03 | 0.82 |

Model selected: Negative Binomial

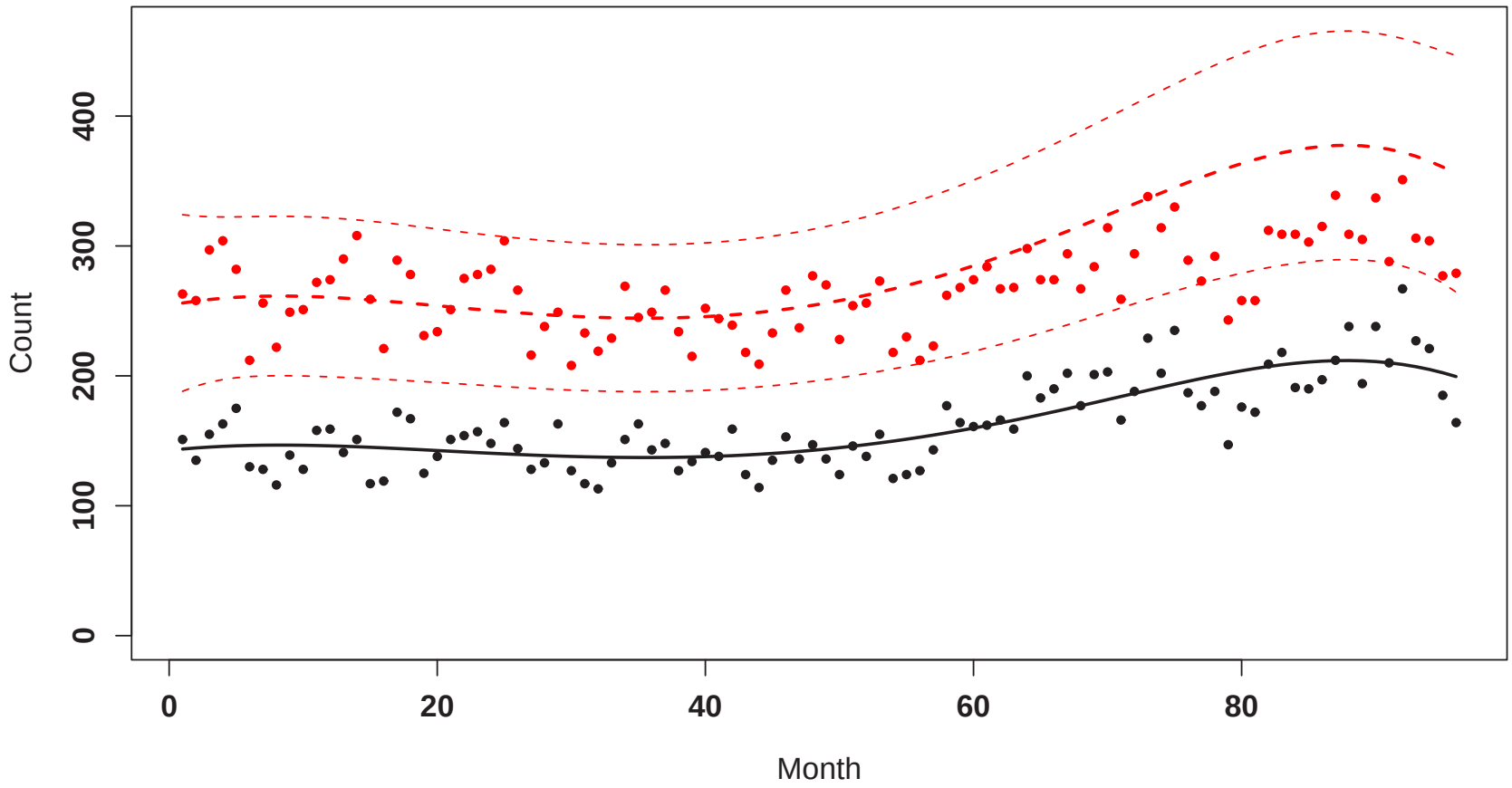
$$\hat{\pi} = 0.55,\ p(D) = 0.63$$

# Heart Attack Data




Estimated mean (solid), estimated total number of heart attacks with confidence interval (dashed)

# Heart Attack Counts + Cause of Death Counts




Estimated mean (black solid), estimated total number of heart attacks with confidence interval (red dashed)

# Over-Reporting

**How to estimate the correct number of cases?**

# Cameron & Trivedi, 1998

Model dependence between the two Bernoulli variables

$E =$ Event and $R =$ Recording

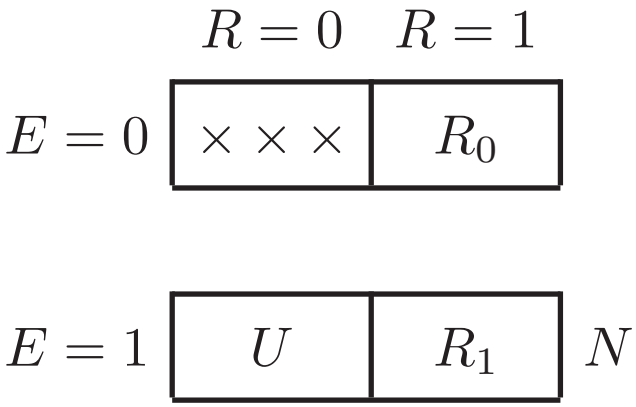
Bivariate Binomial Random Variable

| | $R=0$ | $R=1$ | |
|---|---|---|---|
| $E=0$ | $T_0$ | $O$ | $N_0$ |
| $E=1$ | $U$ | $T_1$ | $N_1$ |
| | $N-Y$ | $Y$ | $N$ |

$T_0$: true not reported
$T_1$: true reported
$O$: over-reported
$U$: under-reported
$Y$: observed count

# Independent Sampling

In Criminology $T_0$ does not exist. Therefore we consider independent sampling.




and

$$Y = R_0 + R_1$$

Specify under-reporting model for $R_1$ and a count model for $R_0$

Model $Y$ by convolution of both

# A Family of Poisson Convolutions

Assume that $R_1$ is generated by under-reporting, i.e.

$$R_1|N,P \sim \text{Binomial}(N,P),$$

and

$$R_0 \sim \text{Poisson}(\alpha)$$

Then

$$p(Y=y=r_0+r_1) = \sum_{j=0}^{r_0} p_0(j)p_1(y-j) = \sum_{j=0}^{r_1} p_0(y-j)p_1(j)$$

with

$$\begin{aligned} E(Y) &= \mu_0+\mu_1 = \alpha+\lambda\pi \\ var(Y) &= \sigma_0^2+\sigma_1^2 = \alpha+var(R_1) \end{aligned}$$

# Negative-Binomial-Poisson Convolution

For $R_1 \sim$ Negative Binomial$(\omega, \pi)$

$$Y \sim \text{Delaporte}(\lambda, \pi, \alpha), \qquad \lambda = \omega/(1-\pi)$$

with

$$p(Y = y = r_1 + r_0) = \frac{(1-\pi)^{\omega}\alpha^{y}\exp(-\alpha)}{\Gamma(\omega)}S$$

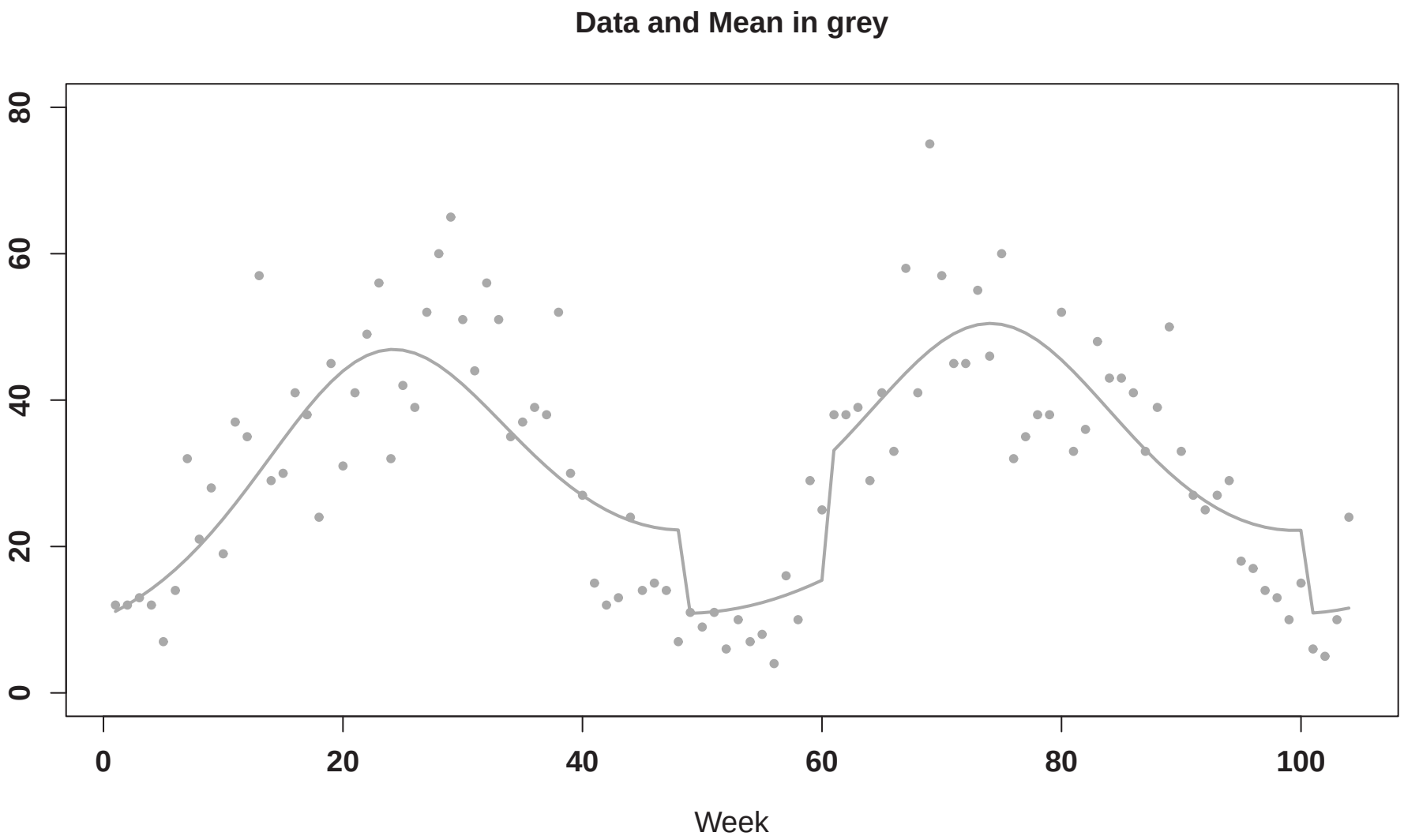
where

$$S = \sum_{j=0}^{r_1} \frac{\Gamma(j+\omega)}{\Gamma(j+1)\Gamma(y-j+1)}\left(\frac{\pi}{\alpha}\right)^{j}$$

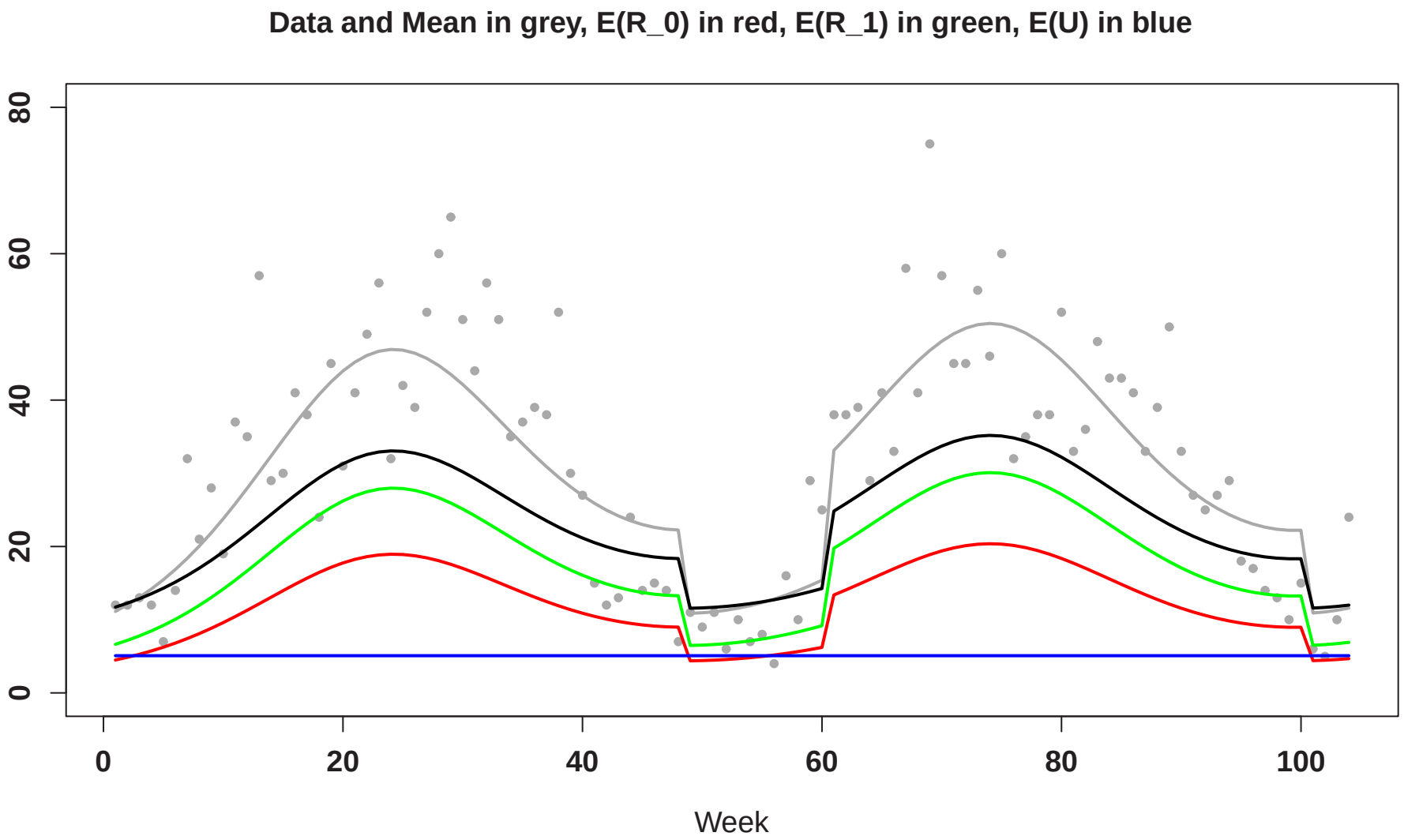
and

$$\begin{aligned} E(Y) &= \mu_0 + \mu_1 = \alpha + \lambda\pi \\ var(Y) &= \sigma_0^2 + \sigma_1^2 = \alpha + \lambda\pi(1-\pi)^{-1} \end{aligned}$$

# Application to bicycle theft data



# Application to bicycle theft data



# Application to bicycle theft data

Averaged Results

| | $R=0$ | $R=1$ | |
|---|---|---|---|
| $E=0$ | $\times\times\times$ | $\hat{E}(R_0)=14.35$ | |
| $E=1$ | $\hat{E}(U)=5.08$ | $\hat{E}(R_1)=21.18$ | $\hat{E}(N)=26.26$ |

and

$$\hat{E}(Y)=\hat{E}(R_0+R_1)=35.53$$

Reporting probability in the under-reporting model: $\hat{\pi}=0.77$

Fraud rate: $14.35/35.53=0.40$

# Conclusion

- highly relevant methodology for wide-spread applications
- models based on Bernoulli sampling
- conditional binomial models suitable for cases when $\mathrm{var}(Y) > \mu$
- estimation relies on ML and Hybrid ML
- non-nested technique for model selection

Limitations

- $\pi \longrightarrow 0$, i.e. Poisson limit
- $\pi \longrightarrow 1$, perfect reporting system

# References


Allcroft, D.J. and Glasby, Ch.A. (2003). A simulation-based method for model evaluation, *Statistical Modelling,* 3, 1-14.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Neubauer, G., Djuraš, G. and Friedl, H. (2011). Models for Underreporting: A Bernoulli Sampling Approach for Reported Counts. *Austrian Journal of Statistics,* 40, 85-92.

Winkelmann, R. (2000). *Econometric Analysis of Count Data.* Berlin: Springer.