

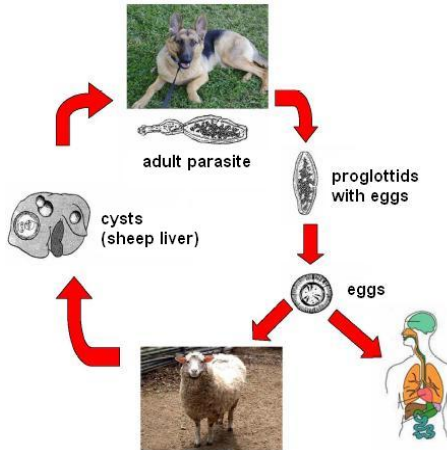
# Hierarchical Poisson models for nonzero count estimation

Dominik Heinzmann

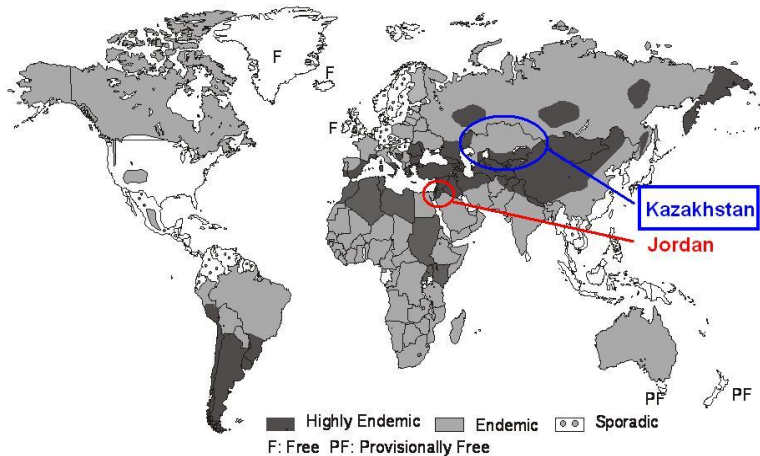
25. September 2007

Institute of Mathematics, University of Zurich  
Institute of Parasitology, University of Zurich

# Life cycle of *E. granulosus* (Dog tapeworm)



# Geographical distribution

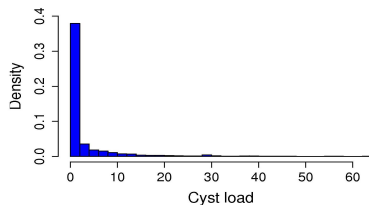


# Overview

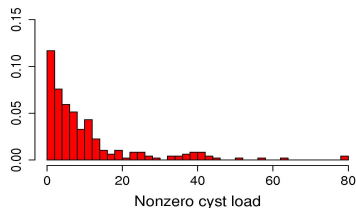
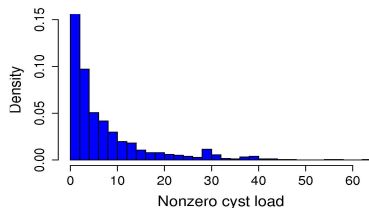
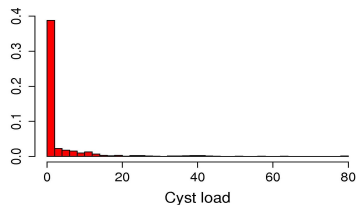
- Two-part conditional model
- Zero-truncated negative binomial (ZTNB)
- Filtered polynomial mass function (FPMF)
- Comparison

# Empirical cyst distribution in sheep

Kazakhstan



Jordan



# Features of *E. granulosus* hydatid cysts:

- Substantial proportion of zeros.
- Remaining nonzero loads are positively skewed.
- Assumptions: Two part infection mechanism
- First part:
  - Mechanism: Contact/ingestion of (infective) fecal clumps
  - Depend on: Dogs behavior, sheep behavior
- Second part:
  - Mechanism: Number of cysts acquired conditional on infectivity of the clump
  - Depend on: Worm load in dog, genetic and/or physiological conditions (e.g. immune response)

# Two-part conditional model

In what follows:  $Y$  is a random variable modeling the cyst load in sheep.

$$\mathbb{P}(Y = 0) = 1 - q$$

$$\mathbb{P}(Y = y) = qf_{ZT}(y) , y = 1, 2, \dots ,$$

where

$q$  : Probability of observing nonzero loads (prevalence).

$f_{ZT}(y)$  zero-truncated probability mass function (pmf).

# Log-likelihood

Given a sample of nonzero loads  $y = (y_1, \dots, y_n)$ ,

$$\begin{aligned}\log L(y) &= \sum_{i=1}^n I(y_i = 0) \log(1 - q) + I(y_i > 0) \log(q) \\ &\quad + \sum_{i=1}^n I(y_i > 0) \log(f_{Z_T}(y_i)) \\ &= \sum_{i=1}^n [I(y_i = 0) \log(1 - q) + (1 - I(y_i = 0)) \log(q)] \\ &\quad + \sum_{i=1}^n I(y_i > 0) \log(f_{Z_T}(y_i)).\end{aligned}$$

$\Rightarrow q$  and  $f_{Z_T}$  can be estimated separately by MLE

# Hierarchical (mixed) Poisson model

Let be given non-negative RV  $Y$  (integer-valued) and  $\Lambda$  (real-valued).

$$Y \sim \text{Pois}(\Lambda), \text{ where } \Lambda \sim h(.).$$

Typical choice: gamma distribution

$$h(\lambda) = h(\lambda; \psi, \xi) = \frac{1}{\xi^\psi \Gamma(\psi)} \lambda^{\psi-1} \exp(-\lambda/\xi),$$

where  $\psi, \xi > 0$ , and hence we obtain the pmf of the NB distribution.

$$\begin{aligned} f(y; \psi, \xi) &= \int_0^\infty g(x|\lambda) h(\lambda; \psi, \xi) d\lambda \\ &= \frac{\Gamma(\psi + y)}{\Gamma(\psi) y!} \left( \frac{1}{1 + \xi} \right)^\psi \left( \frac{\xi}{1 + \xi} \right)^y. \end{aligned}$$

# Truncation

The pmf of  $Y$  conditional on  $Y > 0$  can be written as

$$\begin{aligned}f_{ZT}(y; \psi, \xi) &= f(y; \psi, \xi | y > 0) \\&= \mathbb{P}_{\psi, \xi}(Y = y | Y > 0) \\&= \frac{f(y; \psi, \xi)}{\mathbb{P}_{\psi, \xi}(Y > 0)} \\&= \frac{f(y; \psi, \xi)}{1 - \mathbb{P}_{\psi, \xi}(Y = 0)}.\end{aligned}$$

We will refer to this pmf as zero-truncated negative binomial function (ZTNB).

# Filtered polynomial density estimation

Let  $X$  be a real-valued RV.


Every unknown continuous distribution function  $F(x)$  and the corresponding pdf  $f(x) = F'(x)$  can be estimated as follows (Heinzmann, 2007<sup>1</sup>).

$$F(x) \approx H(m_{2k+1}(x))$$
$$f(x) \approx h(m_{2k+1}(x))p_{2k}(x),$$

where

- $H(\cdot)$  is the filter and  $h(\cdot) = H'(\cdot)$  its derivative.
- $m_{2k+1}(x)$  is a monotonic increasing polynomial of degree  $2k + 1$
- $p_{2k}(x) = m'_{2k+1}(x)$  its derivative and thus a positive polynomial of degree  $2k$ .

---

<sup>1</sup>Heinzmann, D. (2007). A filtered polynomial approach to density estimation, *Computational Statistics*, doi:10.1007/s00180-007-0070-z. 

# Computational aspects: Positive polynomial

- Positive polynomial order 2

$$u_i = 1 - 2a_i x + b_i x^2 = (a_i x - 1)^2 + c_i x^2 \geq 0$$

where  $b_i = a_i^2 + c_i$

- Positive polynomial order  $2k$

$$p_{2k}(x) = \gamma \prod_{j=1}^k u_j = h_0 + h_1 x + \dots + h_{2k} x^{2k}$$

- To ensure positivity

$$c_i = \tilde{c}_i^2 \quad \text{and} \quad \gamma = \exp(\alpha)$$

# Computational aspects: Monotonic polynomial

- Corresponding monotonic increasing polynomial order  $2k + 1$

$$m_{2k+1}(x) = \omega_0 + h_0x + \frac{h_1}{2}x^2 + \dots + \frac{h_{2k}}{2k+1}x^{2k+1}.$$

- Coefficient  $\Theta$  for a monotonic polynomial of degree  $2k + 1$  be written as

$$\Theta_{2k+1} = (\omega_0 \quad \alpha \quad a_1 \quad \tilde{c}_1 \quad a_2 \quad \tilde{c}_2 \dots \quad a_k \quad \tilde{c}_k)^T.$$

- Thus we have a nested polynomial structure which facilitates the optimization.

# Filtered polynomial mass function (FPMF)

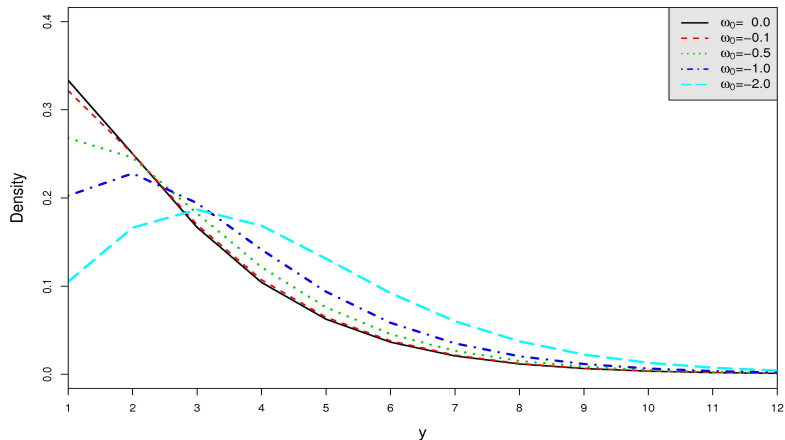
Now, we apply the FPDE approach to the mixture distribution:

$$\begin{aligned} f(y; \psi, \xi) &= \int_{\lambda^*}^{\infty} g(y|\lambda) h(m(\lambda); \psi, \xi) p(\lambda) d\lambda \\ &= \int_{\lambda^*}^{\infty} \frac{\lambda^y}{y!} \exp(-\lambda) \frac{1}{(\xi)^\psi \Gamma(\psi)} m(\lambda)^{\psi-1} \exp\left(\frac{-m(\lambda)}{\xi}\right) p(\lambda) d\lambda, \end{aligned}$$

where  $m(\lambda) = \omega_0 + \sum_{i=1}^{2k+1} h_{i-1} \lambda^i$  and  $\lambda^*$  is such that  $m(\lambda^*) = 0$ .

- $h_0$  can be set to 1 (scaled by  $\xi$ ).
- $m(\lambda) = \lambda$  implies NB.

# Influence FPDE for $m(\lambda) = \omega_0 + \lambda$



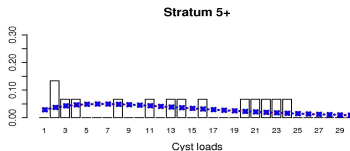
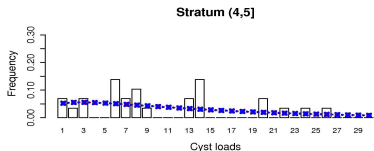
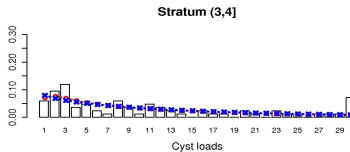
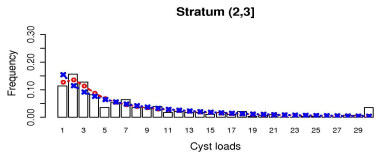
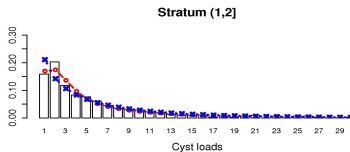
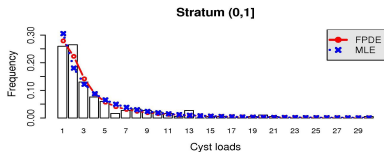
# Example Data

- Sample from Kazakhstan<sup>2</sup> with size 2505.
- Individual reports of age and hydatid cyst burden.
- After slaughtering:
  - ① Counting the cysts.
  - ② Age ascertained by careful examination of its dentition.
- Age-stratification ("average out" progressively acquisition of clumps).

---

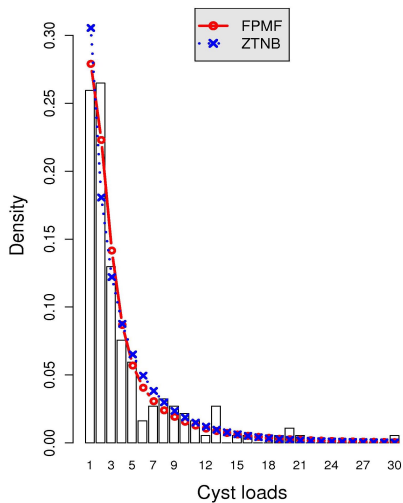
<sup>2</sup>Torgerson, P.R., Shaikenov, B.S., Rysmukhambetova, A.T., Ussenbayev, A.E., Abdybekova, A.M. and Burtisurnov, K.K. (2003), Modelling the transmission dynamics of *Echinococcus granulosus* in sheep and cattle in Kazakhstan. *Vet Parasitol*, **114**, 143-153.

# FPDE Kazakhstan

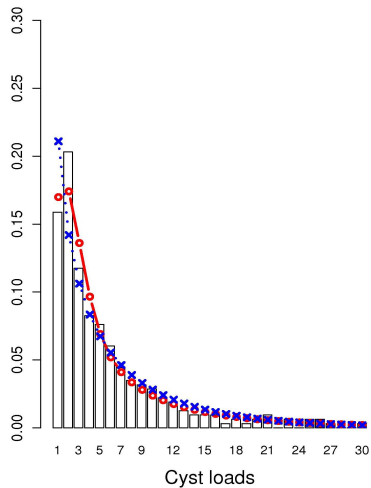


# FPDE Kazakhstan

## Stratum (0,1]



## Stratum (1,2]



# Monte-Carlo based p-value

Let be given estimates

- 1  $\hat{\eta}_1 = (\hat{\psi}, \hat{\xi})$  for the ZTNB approach
- 2  $\hat{\eta}_2 = (\hat{\psi}, \hat{\xi}, \hat{\Theta})$  for the FPMF approach.

To test the null hypothesis  $H_0 : \Theta = \mathbf{0}$ .

- Compute likelihood ratio test statistic of real data

$$\text{LRT}_0 = -2[\log L(\mathbf{y}; \hat{\eta}_2) - \log L(\mathbf{y}; \hat{\eta}_1)]$$

- Generate data sets  $S_k$  ( $k = 1, \dots, 1000$ ) under  $H_0$  of size equal to the real data.
- Calculate ZTNB and the FPMF fits of  $S_k$ .
- Evaluate LRT statistics  $\text{LRT}_k$  ( $k = 1, \dots, 1000$ ).
- Empirical p-value =  $\#\{k \mid \text{LRT}_k > \text{LRT}_0\}/1000$ .

TABLE: Comparison of FPMF<sub>1</sub> (linear) with ZTNB and FPMF<sub>2</sub> (cubic).

Stratum	NLL <sub>FPMF<sub>1</sub></sub>	NLL <sub>ZTNB</sub>	LRT	p-value	Power
(0, 1]	415.816	419.355	7.078	0.018	0.655
(1, 2]	863.349	869.546	12.394	0.012	0.702
(2, 3]	867.750	871.879	8.258	0.016	0.633
(3, 4]	308.311	308.788	0.954	0.170	0.205
(4, 5]	103.9939	103.9944	0.001	0.742	0.054
(5, ∞)	53.611	53.634	0.045	0.660	0.043
		NLL <sub>FPMF<sub>2</sub></sub>	LRT	p-value	Power
(0, 1]		415.516	0.600	0.741	0.044
(1, 2]		863.338	0.022	0.989	0.052
(2, 3]		867.639	0.222	0.895	0.042
(3, 4]		307.999	0.624	0.732	0.047
(4, 5]		103.386	1.216	0.544	0.039
(5, ∞)		53.341	0.584	0.747	0.051

# Goodness-of-fit

Criteria to determine number of classes  $c$  in Pearson's  $\chi^2$  test:

- 1 Homogeneous distribution (approximately equal probabilities in classes). (Good et al., 1970)
- 2 If  $m$  is the sample size, then we assure that  $m \geq 10$ ,  $c \geq 3$  and  $m^2/c \geq 10$ . (Koehler and Larntz, 1980)
- 3 Let the expected number in class  $l$  be  $m\hat{p}_l \geq 10$ .

# Goodness-of-fit

TABLE: Results of the FPMF (index 1) and ZTNB (index 2) approach.

	$c$	$df_1$	$\chi^2_{P,1}$	$p_1$	$df_2$	$\chi^2_{P,2}$	$p_2$
(0, 1]	8	4	7.785	0.100	5	13.305	0.021
(1, 2]	10	6	5.796	0.446	7	14.963	0.036
(2, 3]	11	7	10.188	0.178	8	18.656	0.017
(3, 4]	7	3	7.068	0.069	4	9.166	0.057
(4, 5]	5	1	3.319	0.068	2	3.319	0.190
(5, $\infty$ )	5	1	1.795	0.180	2	1.795	0.408

# Summary

Based on our sample from Kazakhstan, we have shown that

- FPMF performs significantly better than the ZTNB in most of the age strata.
- The FPMF provides (much) better estimates of the mean cyst burden and the corresponding variance than the ZTNB.
- For strata, where the ZTNB has a lack of fit (Person's  $\chi^2$  test), the FPMF does not provide such a lack.
- Similar results have been obtained by applying the approach to a sample from Jordan (size 832).
- The approach will be challenged by new data from Kyrgyzstan.