A semi-bayesian Goodness-of-Fit Tests in Logistic Regression with Sparse Data

Oliver Kuss
Institute of Medical Epidemiology, Biostatistics and Informatics, Medical Faculty
University of Halle-Wittenberg,
Oliver.Kuss@medizin.uni-halle.de

Programme:

1. The Logistic Regression Model
2. Checking Goodness-of-Fit
3. The Problem of Sparse Data
4. Some Solutions
5. A semi-bayesian solution
6. Discussion
7. Literature

1. The Logistic Regression Model

Standard method for the regression of binary responses

Reasons:

- Easy interpretation of parameters as odds-ratios
- One can predict response probabilities
- Software is available
- Valid in prospective and retrospective designs

Notation:

N independent observations **grouped by covariate patterns** (y_i , x_i), i=1, ..., N

x_i : vector of p+1 covariates,

y_i : number of successes, $Y_i \sim B(m_i, \pi_i)$,

m_i : number of trials,

$M=\sum_{i=1}^{N} m_i$: Number of individual observations

Data:

		Response 1	Response 0	
Covariate Pattern	1	Y_1	m_1-Y_1	m_1
	2	Y_2	m_2-Y_2	m_2
	:	:	:	:
	N	Y_N	m_N-Y_N	m_N

Example:

Continuous covariate(s): N=M ($m_i \equiv 1$)

		Response 1	Response 0	
Covariate Pattern	1	1	0	1
	2	0	1	1
	:	:	:	:
	N	1	0	1

Model equation:

$$\log\left(\pi_i \big/ 1-\pi_i\right) = \sum_{j=0}^{p} x_{ij}\beta_j$$

with $\beta_j = (\beta_0,...,\beta_p)$ vector of regression parameters.

Estimate parameters β_j via ML.

2. Checking Goodness-of-Fit

Statistical modelling consists of two steps (Hosmer et al., 1991):
Model building and **Model checking**

Model building: How can I explain the variation in response values in terms of the covariates (*systematical component*)

Model checking: Assess all variation that is not explained by the systematical component by comparing observations and prognoses from the model (*error component*)

Two different kinds of model checking:

1) Consider deviations from prognoses for each single observations (numerically and graphically): $\Rightarrow$ **Residual analysis**
2) Calculate goodness-of-fit measures and assess model fitting by a single number and a statistical test:$\Rightarrow$ **Goodness-of-fit tests**

A classical Goodness-of-Fit test:

Pearson statistic:

$$X^2 = \sum_{i=1}^{N} \frac{\left(y_i - m_i \hat{\pi}_i\right)^2}{m_i \hat{\pi}_i \left(1 - \hat{\pi}_i\right)}$$

Large values of X^2 indicate lack-of-fit

Statistical test: Compare X^2 to a χ^2-distribution with N-p-1 df

3. The Problem of Sparse Data

The χ^2-limiting distribution is only valid for large m_i (N fixed, $m_i \to \infty$ for all i)

Unrealistic with a large number of covariates or with continuous covariates

A disaster:
In the case of $m_i \equiv 1$ there is: $X^2 \approx N$

4. Some Solutions

4.1 Modify limiting distribution

- X^2, D are asymptotically normal under N, $m_i \to \infty$ (Osius/Rojek, 1992; McCullagh, 1986)

4.2 Grouping observations

- Hosmer-Lemeshow test (Hosmer/Lemeshow, 1980)
  Maybe the standard test with sparse data nowadays, but it has some deficiencies (Hosmer et al, 1997, Bertolini et al., 2000)

4.3 Use other tests statistics

- $\mathbf{X_F^2}$ (Farrington, 1996)

$$X_F^2 = X^2 + \sum_{i=1}^{N} \frac{-(1 - 2\hat{\pi}_i)}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)} (y_i - m_i \hat{\pi}_i)$$

5. A semi-bayesian solution

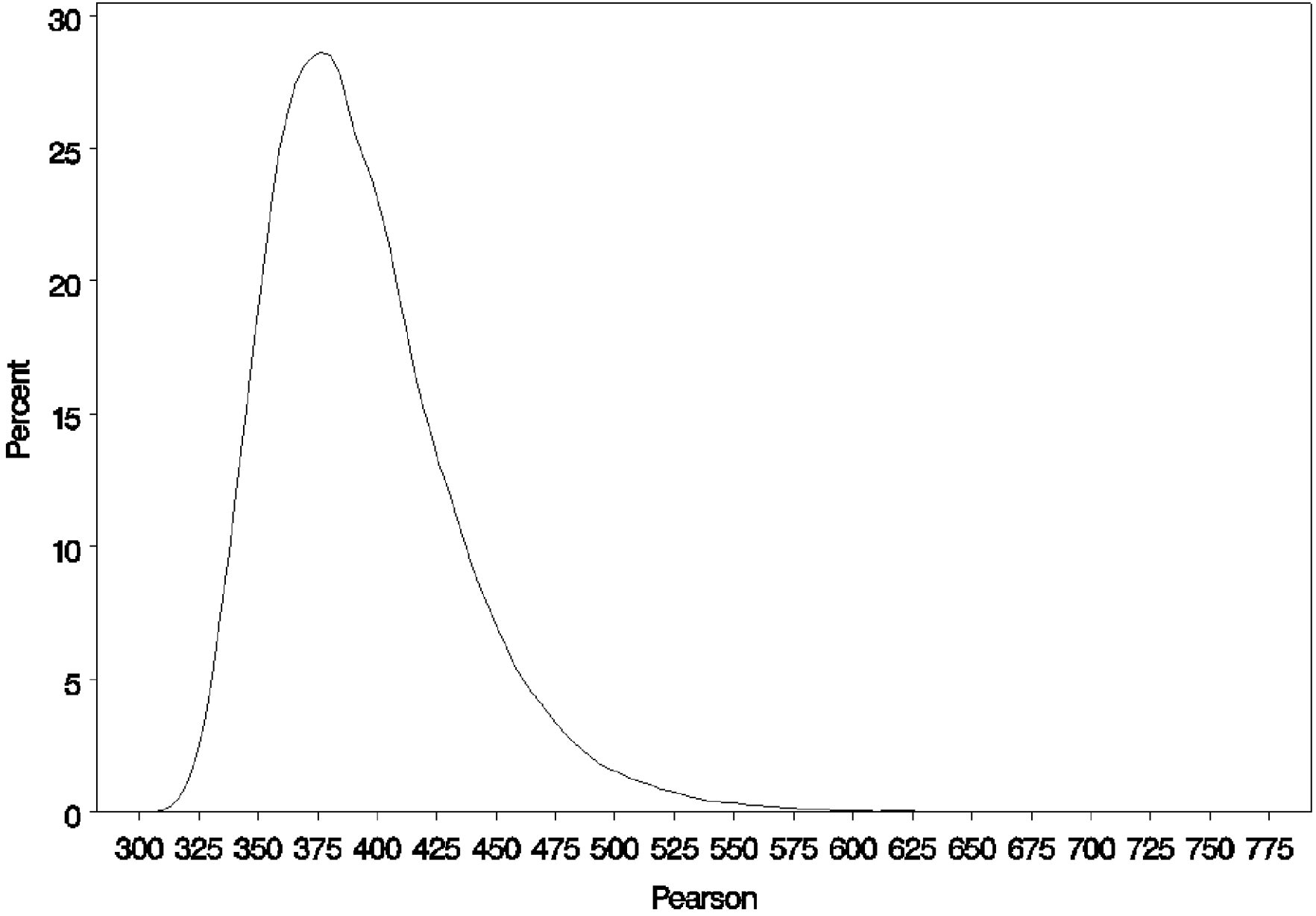
Example:
Occupational hand eczema in hairdresser apprentices,
M=574 (340 „successes“),
Several covariates (p=6): genetical disposition, work hazards, confounders,
N=334,
Distribution of the m_i:

m_i	Frequency
1	205 (61%)
2	68 (20%)
3	35 (11%)
>3	26 (8%)

Step 1:

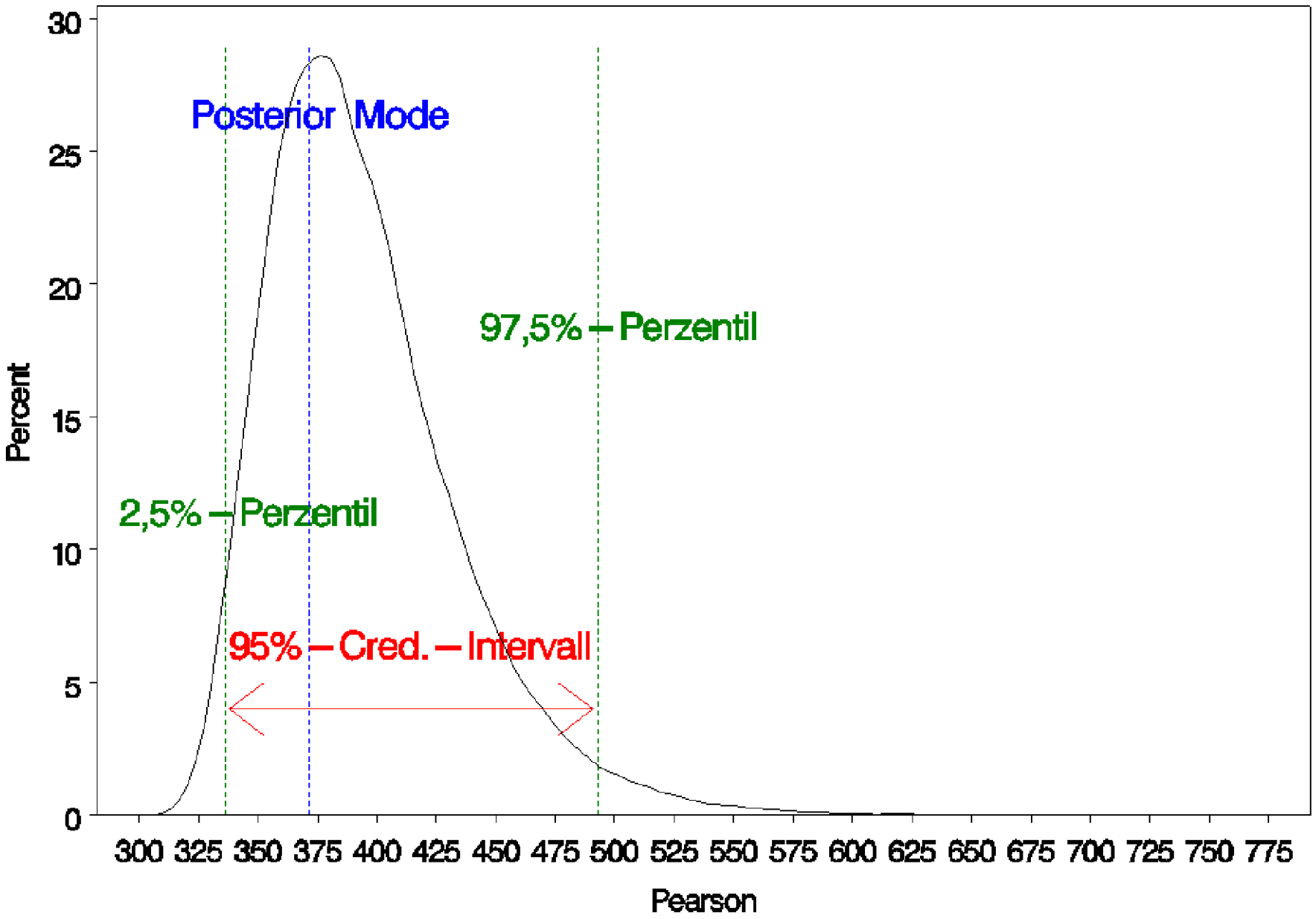
Bayesian estimation of the model gives posterior distribution for all parameters, for all functionals of them, and thus also for the Pearson statistic (see Jackman, 2000)

Thus: Estimate model by MCMC (Gibbs-Sampling, WinBUGS) und get "exact" posterior distribution of X^2


30
25
20
15
10
5
0
Percent
300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775
Pearson

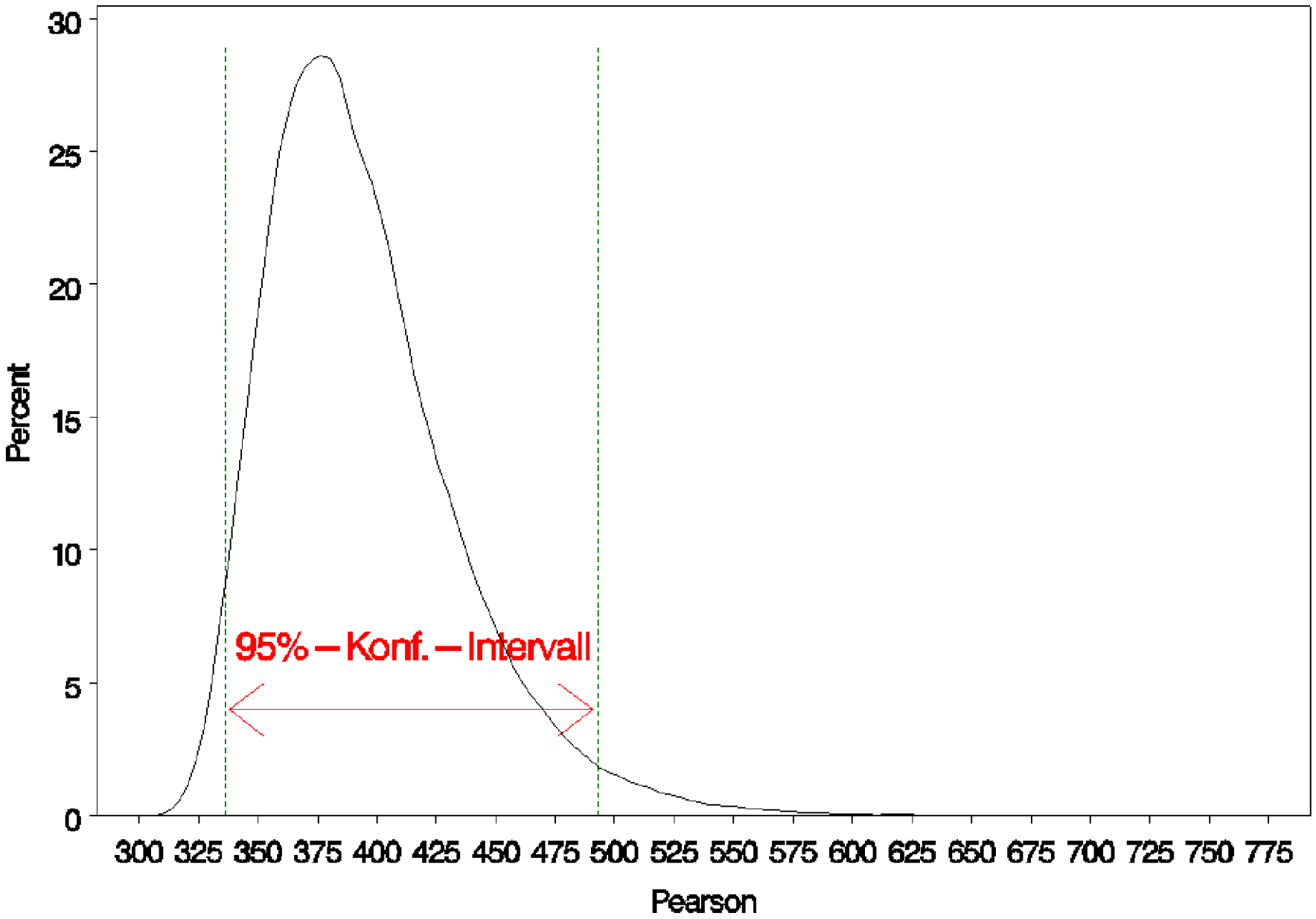
Step 2:

The mode of the posterior gives a new value of the Pearson statistic (X^2_{MCMC}) and by the 2,5%- and 97,5%-percentile a corresponding 95%-credibility interval


Posterior Mode
97,5% – Perzentil
2,5% – Perzentil
95% – Cred. – Intervall
Percent
Pearson
0
5
10
15
20
25
30
300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775

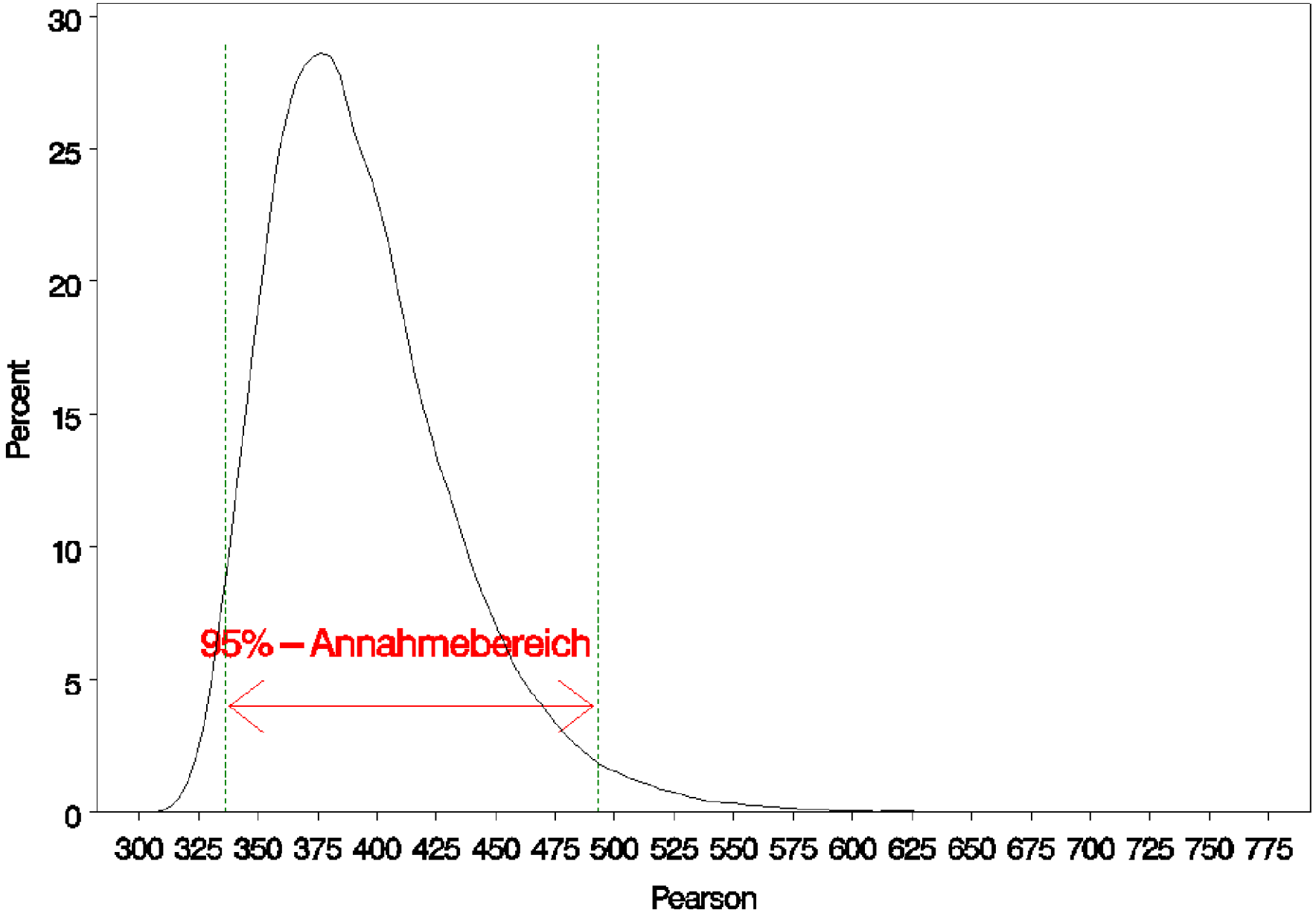
Step 3:

Interpret the 95%-credibility interval as a frequentist 95%-confidence interval (Mossman/Berger, 2001; Carlin/Louis, 2000)


30
25
20
15
10
5
0
Percent
95% – Konf. – Intervall
300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775
Pearson

Step 4:

Identify the 95%-confidence interval as acceptance region of a frequentist statistical test

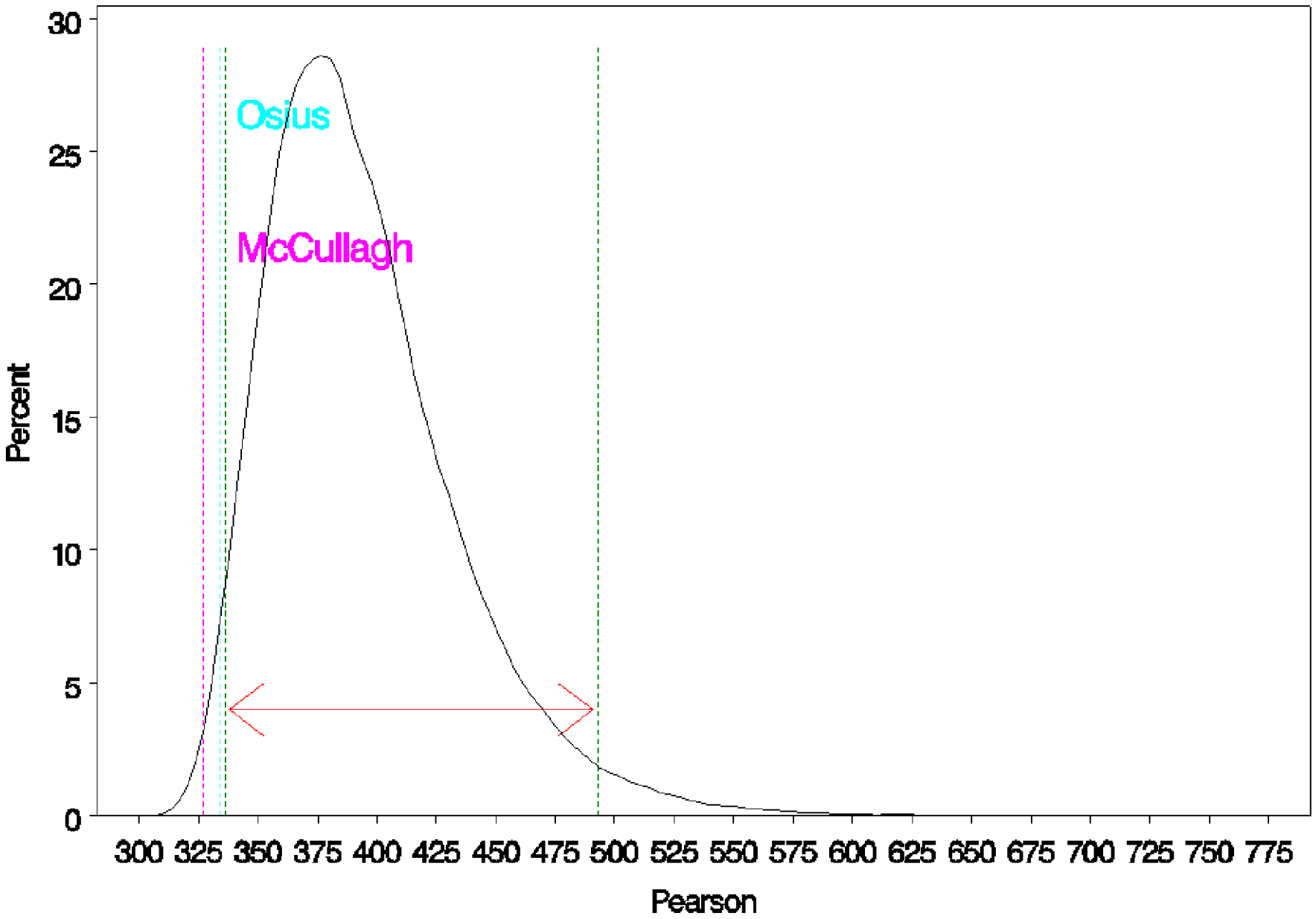

30
25
20
15
10
5
0
Percent
95% – Annahmebereich
300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775
Pearson

Step 5:

Which is the value that indicates the null hypotheses (“The model fits good”)?

Use asymptotic expectations from standard theory:

$E(X^2_O) = N$ (=334; Osius/Rojek, 1992)
$E(X^2_M) = N\text{-}p\text{-}1$ (=327; McCullagh, 1986)


Osius
McCullagh
Percent
0
5
10
15
20
25
30
300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775
Pearson

Summary:

Value of test statistics:

X^2	= 369.25
X^2_{MCMC} (Posterior Mode)	= 371.30
2,5%-percentile	= 335.90
97,5%-percentile	= 493.00

p-values:

X^2	0,053
X^2_{O}	0,044
X^2_{M}	0,031
X^2_{MCMCO}	0,038
X^2_{MCMCM}	0,012

6. Discussion

- The results of the semi-bayesian test seem sensible, maybe better than the standard Pearson test (Kuss, 2002)

- To do:
  Check the MCMC-algorithm (convergence, autocorrelation)
  Check the principle$\rightarrow$ simulation

- The semi-bayesian idea seems to be applicable to other measures in logistic regression (R^2, c, Somer's D etc.) or even to other statistical models

7. Literature

- Bertolini G et al. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *J Epidem Biostat*, 5:251-253, 2000.
- Carlin BP, Louis CA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, 2000.
- Farrington CP. On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data. *J R Statist Soc B,* 58:349-360, 1996.
- Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Statist - Theor Meth*, 9:1043-1069, 1980.
- Hosmer DW et al. A comparison of goodness-of-fit tests for the logistic regression model. *SiM,* 16:965-980, 1997.
- Jackman S. Estimation and Inference Are Missing Data Problems: Unifying Social Science via Bayesian Simulation. *Political Analysis*, 8:307-332, 2000.
- Kuss O. Global Goodness-of-Fit Tests in Logistic Regression with Sparse Data. *Statistics in Medicine*, 21:3789-3801, 2002.
- Mossman D, Berger JO. Intervals for posttest probabilities: a comparison of 5 methods. *Medical Decision Making*, 21:498-507, 2001.
- McCullagh P. On the Asymptotic Distribution of Pearson´s Statistic in Linear Exponential-Family Models. *Int Stat Rev*, 53:61-67, 1985.
- Osius G, Rojek D. Normal Goodness-of-Fit Tests for Multinomial Models With Large Degrees of Freedom. *JASA,* 87:1145-1152, 1992.