# How many cases do we miss when we screen human populations for disease?

Dankmar Böhning

Department of Mathematics and Statistics
School of Mathematical and Physical Sciences
University of Reading

September 6, 2011

the problem

issues with existing approaches

using regression to estimate a probability distribution

predicting the missing cases

discussion

# a motivating problem

## the idea of screening

- human populations are screened for specific diseases
- the aim is to detect the disease early when it is easier to treat and cure
- an example is screening for bowel cancer

# a motivating problem

## screening for bowel cancer

- bowel cancer can develop without any early warning signs
- a test called a FOBT can detect small amounts of blood in the bowel motion
- this might be indicative for a problem such as cancer but also something else such as polyps or nothing

## screening study on bowel cancer in Sydney

- from 1984 onwards about 50000 subjects were screened for bowel cancer in Sydney
- the screening test comprises a sequence of **6 binary diagnostic tests**
- all are self-administered on successive days
- each records the absence or presence of blood in faeces
- details in Lloyd and Frommer (2008)

## screening study on bowel cancer in Sydney

- out of exactly 49927, 46553 tested negatively on all six tests and were not further assessed (diagnosis: not diseased)
- out of the other 3374 subjects who tested positively **at least once**, 3106 were examined and their true disease status determined
- results could be: **healthy, polyps, cancer**
- the other 268 subjects who tested positively were lost to the study

## screening study results in the following table:

Table: *screening of 49927 subjects in Sydney for bowel cancer with partial verification of disease status*

| status | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| healthy | ? | 1123 | 264 | 103 | 35 | 25 | 17 |
| polyps | ? | 772 | 245 | 108 | 72 | 45 | 69 |
| cancer | ? | 46 | 27 | 26 | 33 | 39 | 57 |
| marginal | 46553 | 1941 | 536 | 237 | 140 | 109 | 143 |

## the problem in statistical terms

- $X$ number of positive tests per subject
- $p_x = P(X = x)$ is the probability that exactly $x$ tests are positive for $x = 0, 1, ..., 6$
- $p_0, p_1, p_2, ..., p_6$
- observed is $f_1, ..., f_6$, but $f_0$ is **unobserved**

## the problem in statistical terms

- let $N = f_0 + n = f_0 + f_1 + f_2 + ... + f_6$
- then

$$E(n) = N(1 - p_0)$$

so that a moment estimator gives

- 

$$\hat{N} = n/(1 - p_0)$$

if $p_0$ would be **known**

## how to estimate $p_0$?

1. completely nonparametrically?
2. parametrically?
3. semi-parametrically?

## completely nonparametrically?

- interest in: $p_0, p_1, p_2, ..., p_6$
- observed is $f_1, ..., f_6$, but $f_0$ is **unobserved**
- so that only an estimate of the **zero-truncated distribution**

$$f_j/n \to p_j/(1-p_0)$$

is available which carries **no information** on $p_0$

## how to estimate $p_0$?

1. completely nonparametrically?
2. parametrically?
3. semi-parametrically?

## parametrically?

- $X$ number of positive tests per subject out of $m = 6$
- one could think of the **binomial distribution**

$$p_x = P(X = x) = \binom{m}{x} \theta^x (1-\theta)^{m-x}$$

where $m = 6$ is the number of tests per subject

- with $p_0 = (1-\theta)^m$ the Horvitz-Thompson estimator is

$$\hat{N} = \frac{n}{1-\hat{p}_0} = \frac{n}{1-(1-\hat{\theta})^m}$$

## parametrically?

- clearly possible
- but simple parametric models like the binomial are seldom appropriate
- and certainly **not appropriate** in the situation we have here
- because ...

## a property of the binomial

- consider ratios

$$\frac{p_{x+1}}{p_x} = \frac{\binom{m}{x+1}\theta^{x+1}(1-\theta)^{m-x-1}}{\binom{m}{x}\theta^{x}(1-\theta)^{m-x}}$$

- 

$$= \frac{m!x!(m-x)!}{m!(x+1)!(m-x-1)!}\frac{\theta^{x+1}(1-\theta)^{m-x-1}}{\theta^{x}(1-\theta)^{m-x}} = \frac{m-x}{x+1}\frac{\theta}{1-\theta}$$

- **hence**

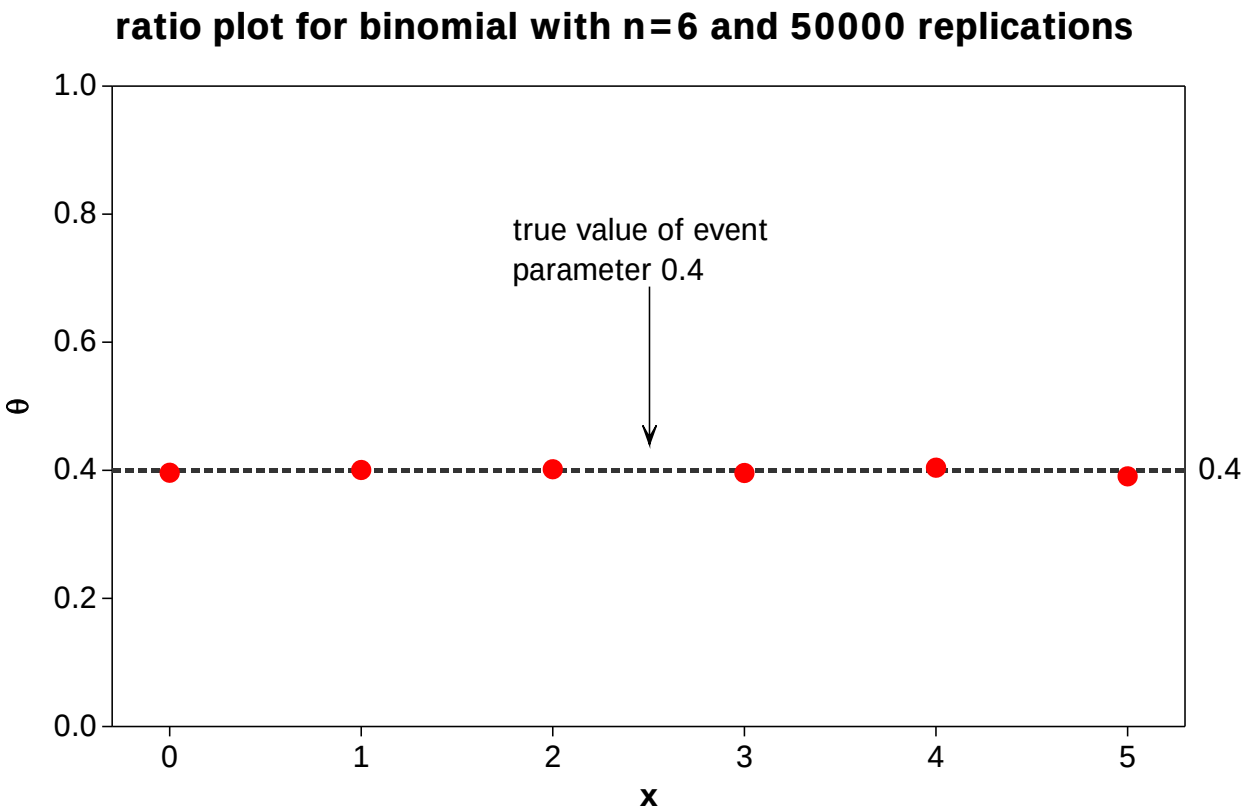
$$a_x\frac{p_{x+1}}{p_x} = \frac{x+1}{m-x}\frac{p_{x+1}}{p_x} = \frac{\theta}{1-\theta}$$

## a diagnostic device for the binomial

- **hence** $x \longrightarrow a_x \frac{p_{x+1}}{p_x}$ is a **horizontal line**
- estimate $a_x \frac{p_{x+1}}{p_x}$ by

$$a_x \frac{f_{x+1}/N}{f_x/N} = a_x \frac{f_{x+1}}{f_x}$$

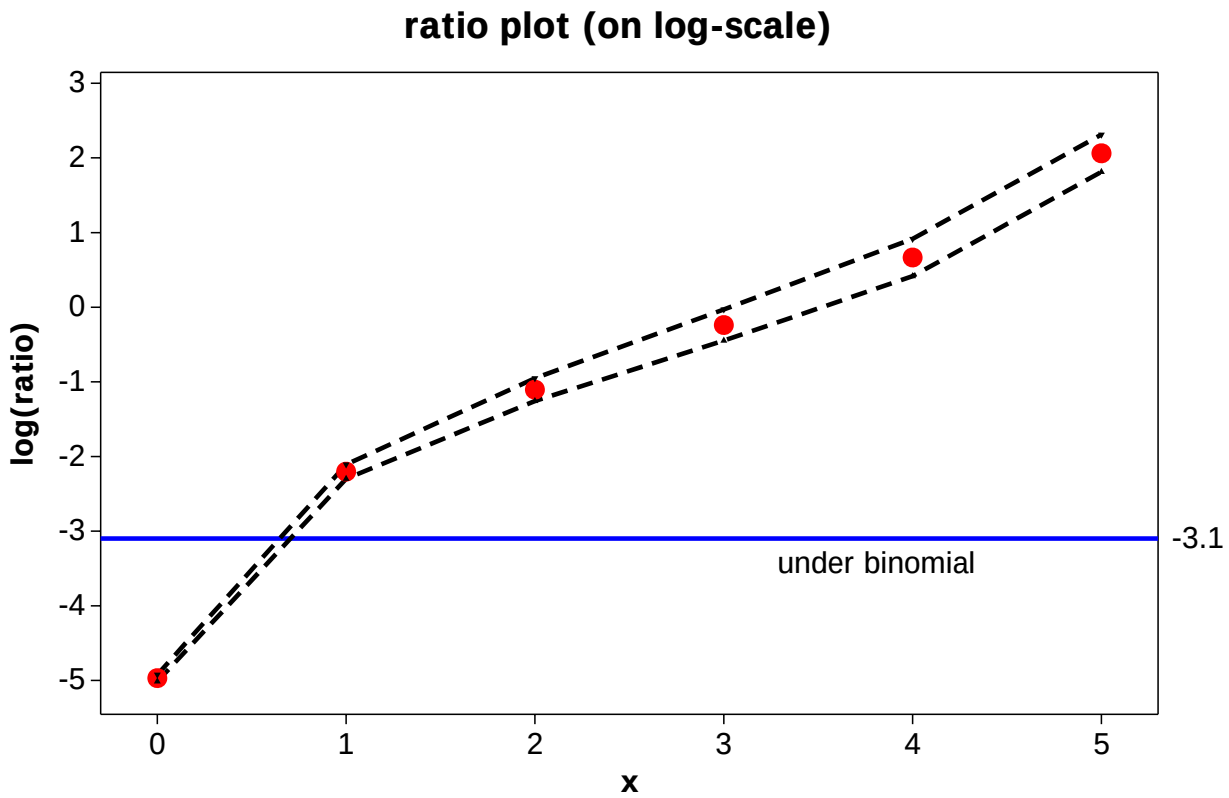
  where $f_x$ is the frequency of count $x$ and $N = f_0 + f_1 + ... + f_m$
- $x \longrightarrow a_x \frac{f_{x+1}}{f_x}$ is a diagnostic device for the binomial and is called the **ratio plot**


ratio plot for binomial with n=6 and 50000 replications
1.0
0.8
0.6
0.4
0.2
0.0
θ
true value of event
parameter 0.4
0.4
0
1
2
3
4
5
x

## application to Sydney screening study:

Table: *screening of 49927 subjects in Sydney for bowel cancer with partial verification of disease status*

| status | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| healthy | ? | 1123 | 264 | 103 | 35 | 25 | 17 |
| polyps | ? | 772 | 245 | 108 | 72 | 45 | 69 |
| cancer | ? | 46 | 27 | 26 | 33 | 39 | 57 |
| marginal | 46553 | 1941 | 536 | 237 | 140 | 109 | 143 |


ratio plot (on log-scale)
log(ratio)
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5
x
under binomial
-3.1

## how to estimate $p_0$?

1. completely nonparametrically?
2. parametrically?
3. semi-parametrically?

## semi-parametrically?

- consider the nonparametric mixture

$$p_x = \int_0^1 \binom{m}{x} \theta^x (1-\theta)^{m-x} g(\theta) d\theta$$

- where $g(\theta)$ is some arbitrary mixing density

## semi-parametrically?

- it was expected that this approach would lead to some great flexibility and would give consistent estimators of $N$ for a wide class of situations:
- $$\hat{N} = \frac{n}{1 - \int_0^1 (1-\theta)^m \, \hat{g}(\theta) d\theta}$$

  where $\hat{g}(\theta)$ would be found on the basis of $f_1, ..., f_m$
- maximizing the so-called zero-truncated mixture likelihood (Böhning and Schön 2005)

## semi-parametrically?

- unfortunately, some things are too beautiful to be true
- Link (2003) showed a **lack of identifiability** of the approach

# Example by Link (2003) on lack of identifiability

## under binomial mixture:

$$p_j = \int_0^1 \binom{4}{j} \theta^j (1-\theta)^{4-j} \, g(\theta) d\theta$$

$j = 0, 1, 2, 3, 4.$

## two mixing distributions:

- uniform $g(\theta) \sim U(a, b)$ with $a = 0.026$ and $b = 0.80$
- discrete two-component mixture
  $g(\theta) \sim 0.576421 \times \delta_{0.286245} + 0.423579 \times \delta_{0.676474}$

# the following table from Link (2003)

Table: *untruncated and truncated count distributions*

| model | probability | count $j$ 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| **uniform** | $p_j$ | 0.227 | 0.255 | 0.243 | 0.190 | 0.085 |
| | $p_j/(1-p_0)$ | - | 0.329 | 0.315 | 0.246 | 0.110 |
| **2 pt. mixture** | $p_j$ | 0.154 | 0.279 | 0.266 | 0.208 | 0.093 |
| | $p_j/(1-p_0)$ | - | 0.329 | 0.315 | 0.246 | 0.110 |

# Consequences of lack of identifiability

- suppose $n = 100$ observed
- using uniform: $\hat{N} = n/0.227 = 440$
- using 2 point mixture: $\hat{N} = n/0.154 = 650$
- very **different values**, but both distributions are indistinguishable as truncated, observable distributions

## ratio plot: evidence for binomial mixture

- evidently, the semi-parametric class is too large
- but how to find a more appropriate class of models?
- we suggest a **regression approach**
- starting point is once more **the ratio plot**

## ratio plot: evidence for binomial mixture

- because: if

$$p_x = \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta) d\theta$$

- then

$$a_0 \frac{p_1}{p_0} \leq a_1 \frac{p_2}{p_1} \leq a_2 \frac{p_3}{p_2} \leq \ldots$$

(Böhning, Baksh, Lerdsruwansri, Gallagher *JCGS* 2011)

- hence **ratio plot is monotone increasing**

## from the ratio plot to a regression problem

$$x \to a_x \frac{p_{x+1}}{p_x} := r_x$$

monotone increasing suggests to look at a **generalised linear model**

$$a_x \frac{p_{x+1}}{p_x} = g^{-1}(\beta' \mathbf{x})$$

where $\mathbf{x}$ is a vector containing several functions of $x$ and $g^{-1}(.)$ is the link-function

**Example:**

$$a_x \frac{p_{x+1}}{p_x} = g^{-1}(\beta_0 + \beta_1 x) = \exp(\beta_0 + \beta_1 x)$$

## every distribution is a regression model

look at the **binomial distribution**

$$a_x \frac{p_{x+1}}{p_x} = \frac{\theta}{1-\theta} = \beta_0$$

or

$$x \rightarrow a_x \frac{p_{x+1}}{p_x} = \beta_0$$

## every distribution is a regression model

look at the **beta-binomial distribution**

$$p_x = \int_0^1 \binom{m}{x} \theta^x (1-\theta)^{m-x} \overbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}}^{g(\theta)} \, d\theta$$

$$= \binom{m}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(m-x+\beta)}{\Gamma(m+\alpha+\beta)}$$

## every distribution is a regression model

leads to

$$x \rightarrow a_x \frac{p_{x+1}}{p_x} = \frac{\Gamma(x+1+\alpha)\Gamma(m-x-1+\beta)}{\Gamma(x+\alpha)\Gamma(m-x+\beta)} = \frac{x+\alpha}{m-x-1+\beta}$$

## every distribution is a regression model

1. every count distribution leads via the ratio plot to a regression model
2. but does **every regression model** lead to a count distribution (under regularity assumptions)?

## every regression model is a distribution

look at the **regression model**

$$a_x \frac{p_{x+1}}{p_x} = g^{-1}(\beta' \mathbf{x})$$

or

$$\frac{p_{x+1}}{p_x} = g^{-1}(\beta' \mathbf{x})/a_x := r_x$$

or

$$p_{x+1} = r_x p_x$$

for $x = 0, 1, ..., n-1$

## Theorem

*Let $r_x > 0$ be given for $x = 0, 1, ..., m-1$. Then there exists a unique count distribution $p_x$ for $x = 0, ..., m$ with the properties*

1. $$p_{x+1} = r_x p_x$$

   *for* $x = 0, 1, ..., m-1$

2. $$p_0 = \frac{1}{1 + r_0 + r_0 r_1 + ... + \prod_{x=0}^{m-1} r_x}$$

**Proof:**

Let $r_x > 0$ be given. Then, using 1.

$$1 = p_0 + p_1 + ... + p_n = p_0 + p_0 r_0 + p_0 r_0 r_1 + ... + p_0 \prod_{x=0}^{n-1} r_x$$

$$= p_0(1 + r_0 + r_0 r_1 + ... + \prod_{x=0}^{n-1} r_x)$$

has to be satisfied. Ultimately, we get

$$p_0 = \frac{1}{1 + r_0 + r_0 r_1 + ... + \prod_{x=0}^{n-1} r_x}$$

## from the ratio plot to a regression problem

an estimate for $a_x \frac{p_{x+1}}{p_x}$ is

$$a_x \frac{f_{x+1}}{f_x}$$

so that ultimately **the model of interest**

$$g(a_x \frac{f_{x+1}}{f_x}) = \beta' \mathbf{x} + \epsilon_{\mathbf{x}},$$

where error $\epsilon_{\mathbf{x}}$ has potentially non-diagonal covariance matrix $\Sigma$

## model fitting

fit the model (in $\beta$)

$$g(a_x \frac{f_{x+1}}{f_x}) = \beta' \mathbf{x} + \epsilon_{\mathbf{x}},$$

for $x = 1, 2, ..., n-1$ using **weighted least squares** and finding **fitted values**

$$a_x \frac{\hat{f}_{x+1}}{\hat{f}_x} = g^{-1}(\hat{\beta}' \mathbf{x}),$$

for $x = 0, 1, 2, ..., n-1$

## finding the fitted distribution from the fitted regression model

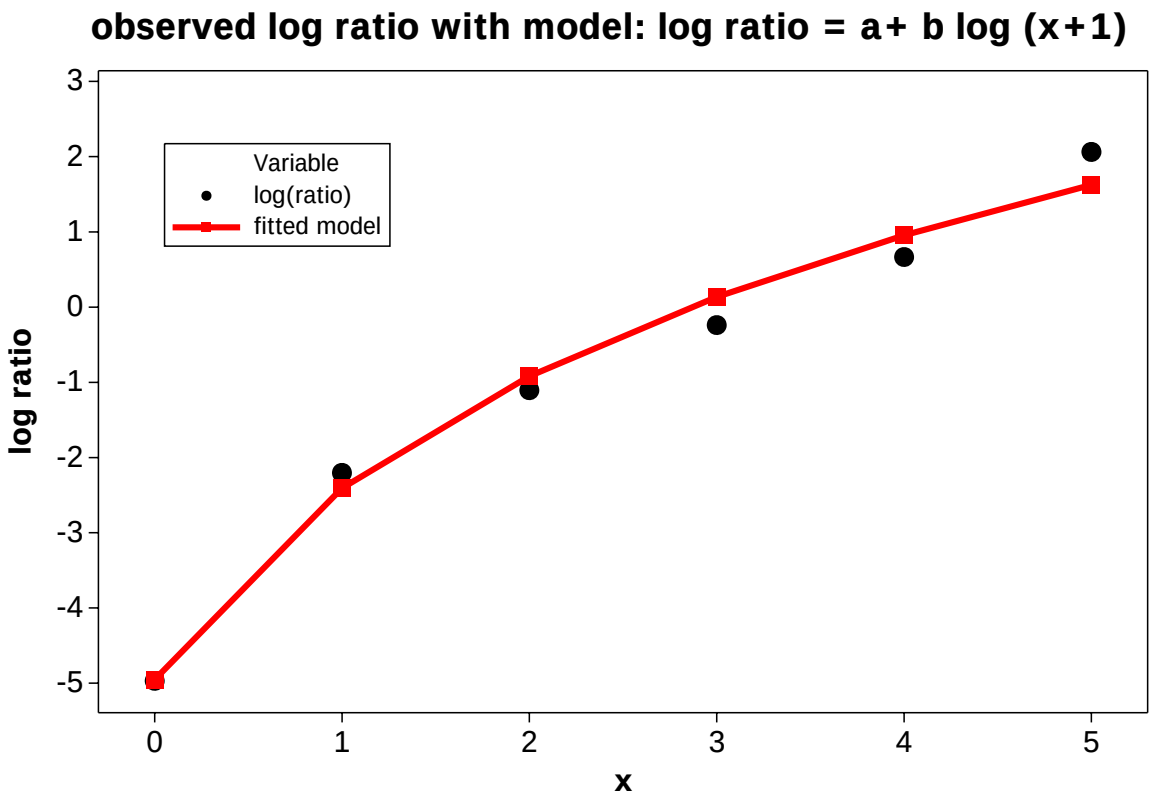
1. $$\hat{r}_x = g^{-1}(\hat{\beta}'\mathbf{x})/\mathbf{a_x}$$

2. $$\hat{p}_{x+1} = \hat{r}_x \hat{p}_x$$

   for $x = 0, 1, ..., m-1$

3. $$\hat{p}_0 = \frac{1}{1 + \hat{r}_0 + \hat{r}_0\hat{r}_1 + ... + \prod_{x=0}^{m-1} \hat{r}_x}$$

# reasonable model: $\log(a_x \frac{f_{x+1}}{f_x}) = \alpha + \beta \log(x+1) + \epsilon_x$



## the log-linear model and GOF

consider the **log-linear model**

$$\log\left(a_x \frac{p_{x+1}}{p_x}\right) = \alpha + \beta \log(x+1)$$

## competitors:

the **beta-binomial model**

$$p_x = \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \overbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}}^{g(\theta)} d\theta$$

$$= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}$$

## competitors:

the **discrete mixture model**

$$p_x = \sum_{j=1}^{J} \binom{n}{x} \theta_j^x (1-\theta_j)^{n-x} w_j$$

where $w_1, ..., w_J \geq 0$ and $w_1 + ... + w_J = 1$

## identifiable mixture models

- Mix2: $J = 2$
- ZIMix2: $J = 3$, but $\theta_1 = 0$
  (zero-inflation model)

# GOF for various models

Table: *Observed and fitted frequencies of the* **marginal** *Sydney screening data with various models fitted: log-linear, binomial model, beta-binomial, two-component mixture, and zero-inflated mixture with two free components*

| model | 0 | 1 | 2 | 3 | 4 | 5 | 6 | $\chi^2$ |
|---|---|---|---|---|---|---|---|---|
| binomial | 44236.6 | 5164.6 | 251.2 | 6.5 | 0.1 | 0.0 | 0.0 | $> 10^9$ |
| beta-bin. | 46842.8 | 1403.9 | 633.2 | 363.0 | 221.5 | 131.1 | 63.6 | 398.9 |
| Mix2 | 46718.9 | 1804.1 | 311.8 | 388.1 | 295.6 | 120.1 | 20.3 | 1053.2 |
| ZIMix2 | 46549.0 | 1881.6 | 639.4 | 160.7 | 137.3 | 175.3 | 100.9 | 97.5 |
| log-linear | 46418.7 | 1968.5 | 443.1 | 235.7 | 202.8 | 211.2 | 178.9 | **96.4** |
| observed | 46553 | 1941 | 536 | 237 | 140 | 109 | 143 | - |

## screening study results in the following table:

Table: *screening of 49927 subjects in Sydney for bowel cancer with partial verification of disease status*

| status | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| healthy | ? | 1123 | 264 | 103 | 35 | 25 | 17 |
| polyps | ? | 772 | 245 | 108 | 72 | 45 | 69 |
| cancer | ? | 46 | 27 | 26 | 33 | 39 | 57 |
| marginal | 46553 | 1941 | 536 | 237 | 140 | 109 | 143 |

## prediction of frequency of zero counts $f_0$

after having fitted the model

$$a_x \frac{\hat{f}_{x+1}}{\hat{f}_x} = g^{-1}(\hat{\beta}' \mathbf{x}),$$

for $x = 1, 2, ..., n-1$: two ways of predicting $f_0$

1. using a Horvitz-Thompson approach
2. using a non-parametric, data-oriented approach

## prediction of frequency of zero counts $f_0$: Way I

after having fitted the model

$$\frac{\hat{f}_{x+1}}{\hat{f}_x} = g^{-1}(\hat{\beta}'\mathbf{x})/\mathbf{a} - \mathbf{x} =: \hat{\mathbf{r}} - \mathbf{x},$$

for $x = 1, 2, ..., n-1$, do the prediction for $x = 0$: $g^{-1}(\hat{\beta}'\mathbf{x_0})$ ($\mathbf{x_0}$ is associated vector for $x = 0$) leading to a fitted distribution via $\hat{p}_{x+1} = \hat{r}_x \hat{p}_x$

$$\hat{p}_0, \hat{p}_1, ..., \hat{p}_m$$

from where the **Horvitz-Thompson estimator** arises:

$$\hat{N} = \frac{n}{1 - \hat{p}_0}$$

## prediction of frequency of zero counts $f_0$: Way II

after having fitted the model

$$a_x \frac{\hat{f}_{x+1}}{\hat{f}_x} = g^{-1}(\hat{\beta}'\mathbf{x}),$$

for $x = 1, 2, ..., n-1$

note that for $x = 0$

$$a_0 \frac{\hat{f}_1}{\hat{f}_0} = g^{-1}(\hat{\beta}'\mathbf{x_0})$$

now, **replacing $\hat{f}_1$ by observed $f_1$** leads to

$$\hat{f}_0 = a_0 \frac{f_1}{g^{-1}(\hat{\beta}'\mathbf{x_0})}$$

## the log-linear model

- $$\log(a_x \frac{f_{x+1}}{f_x}) = \alpha + \beta \log(x+1) + \epsilon_x$$
- $cov(\epsilon) = \Sigma$ not necessarily diagonal
- weighted least squares of $\alpha$ and $\beta$ easily available so that

$$\log(a_0 \frac{\hat{f}_1}{\hat{f}_0}) = \hat{\alpha} + \hat{\beta} \log(1) = \hat{\alpha}$$

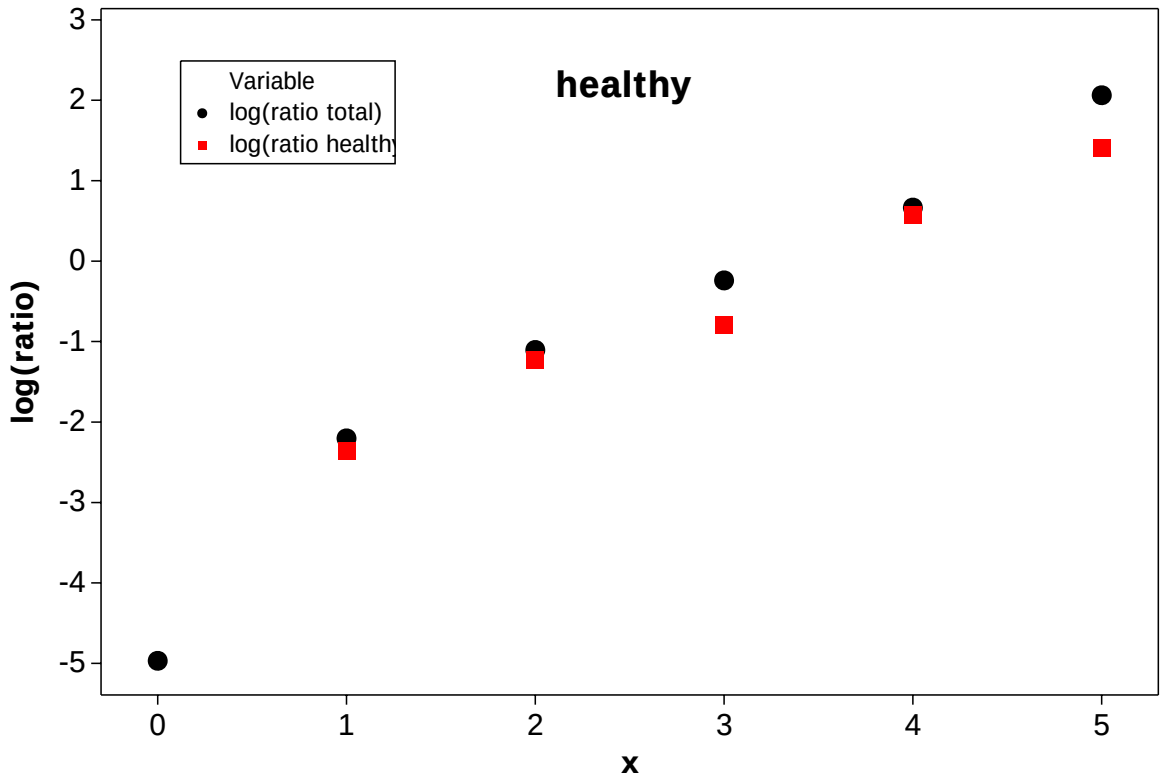
- and will provide **estimate of the missing frequency of zeros**

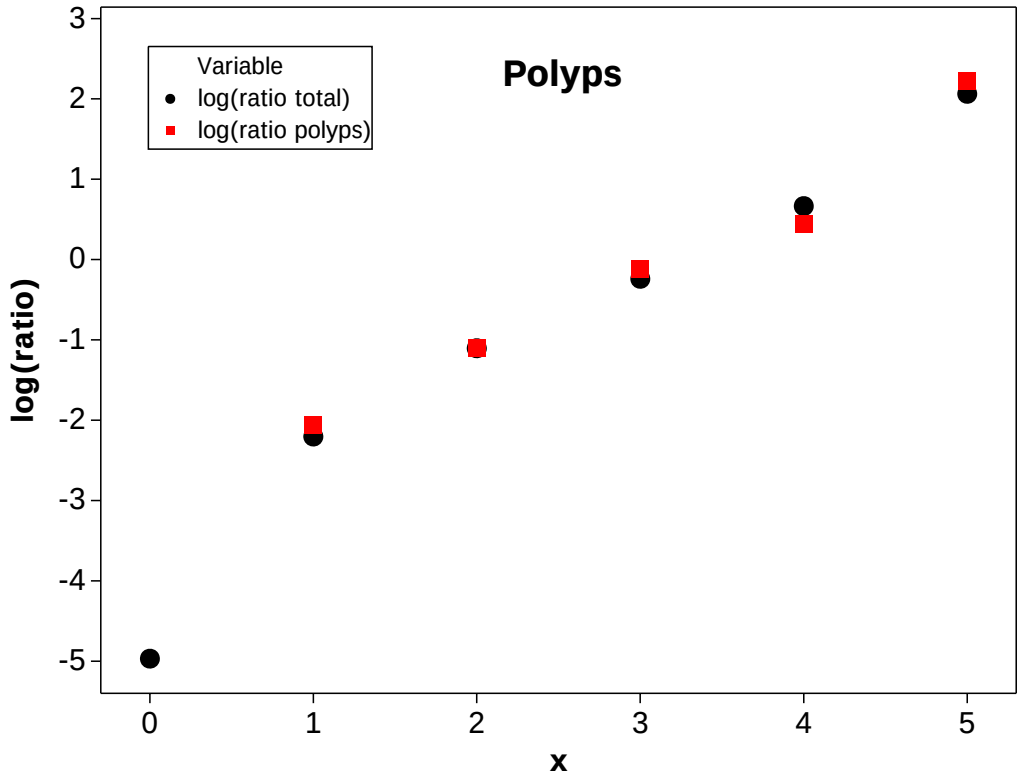
$$\hat{f}_0 = a_0 f_1 \exp(-\hat{\alpha})$$

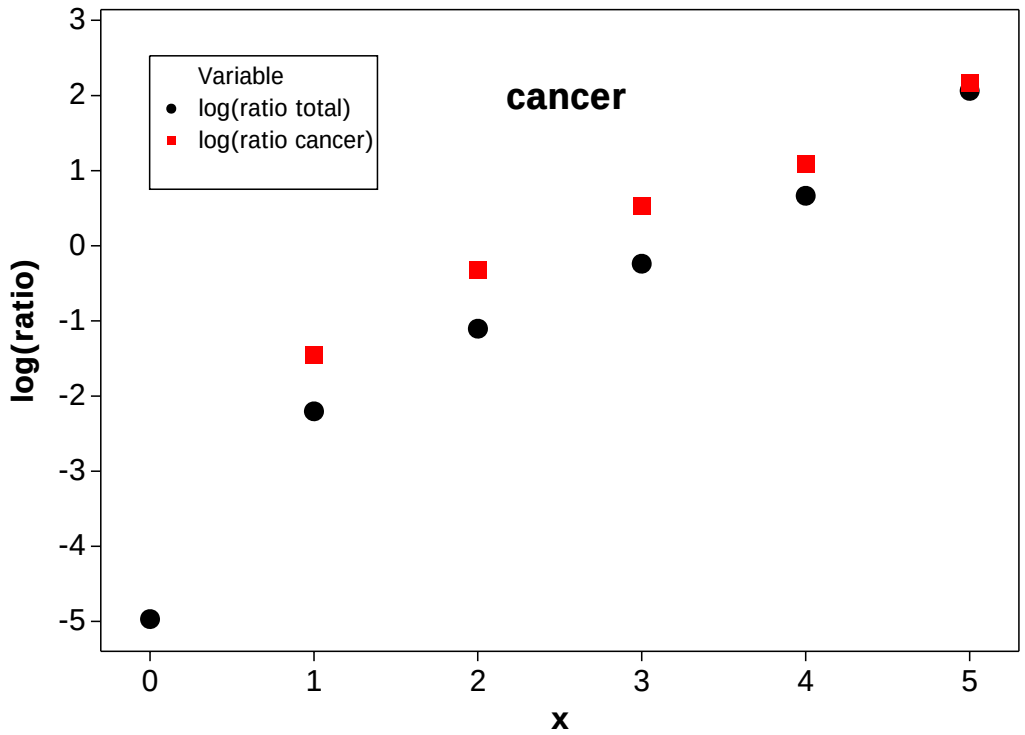
## the model

but is the model developed for the **marginal** frequencies appropriate for the frequencies of the **partially classified** subpopulations:

- with no disease
- with polyps
- with cancer


healthy
Variable
log(ratio total)
log(ratio healthy
log(ratio)
x
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5

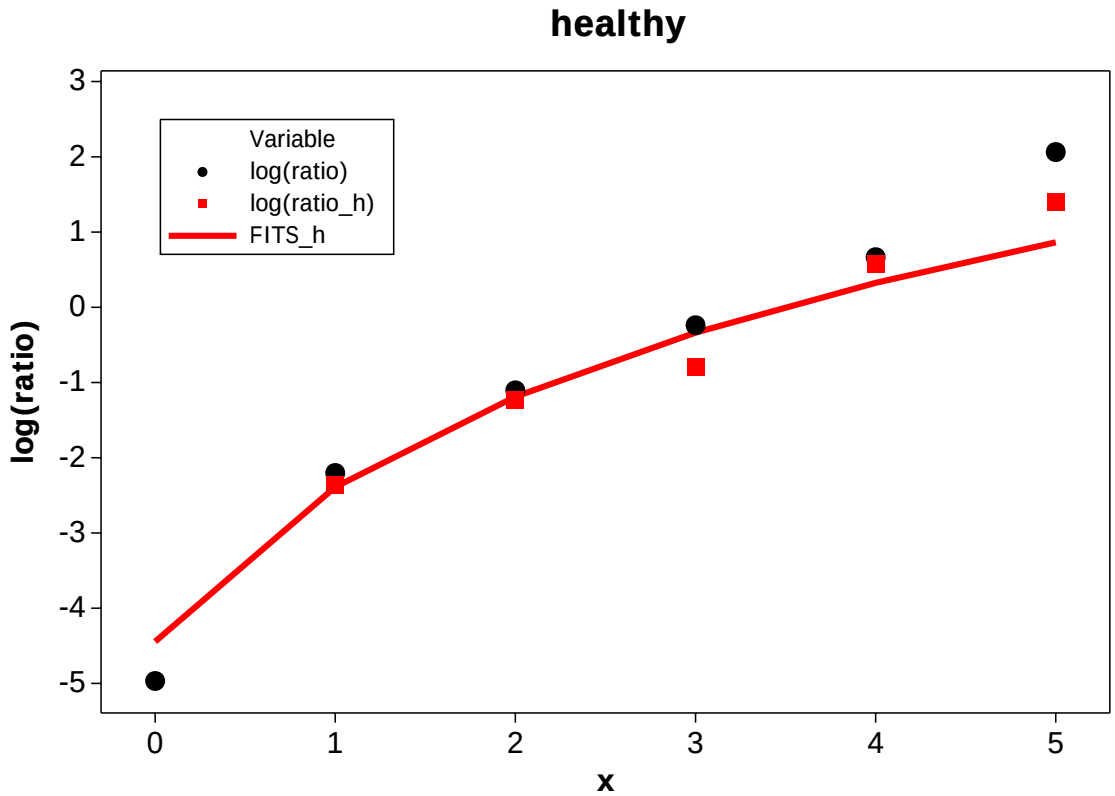

Polyps
Variable
log(ratio total)
log(ratio polyps)
log(ratio)
x
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5

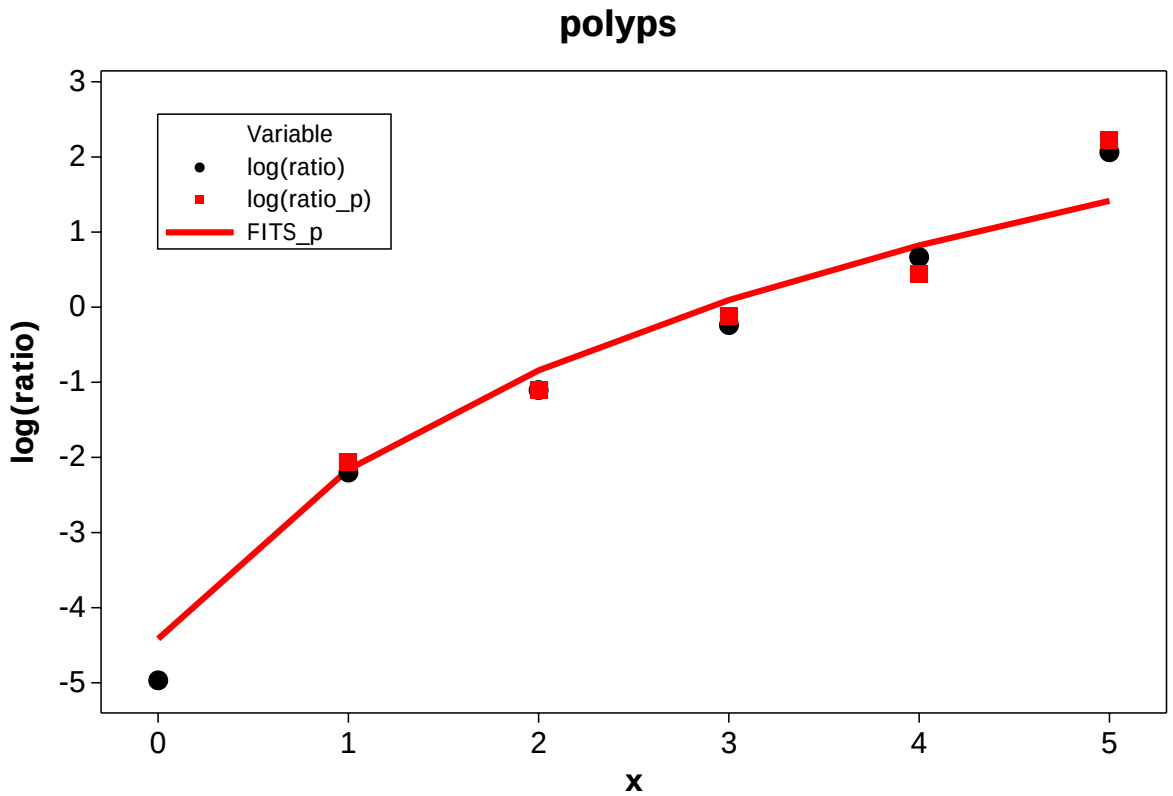

cancer
Variable
log(ratio total)
log(ratio cancer)
log(ratio)
x
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5

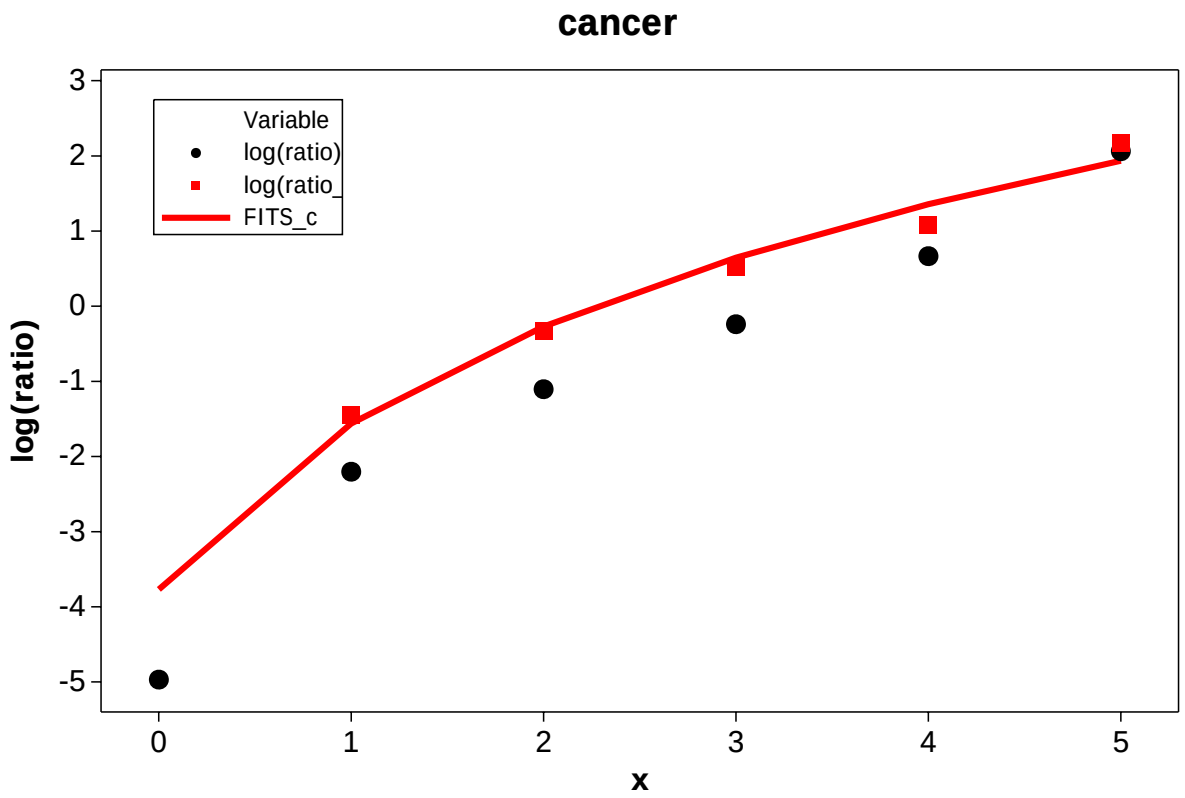
## the prediction

all is needed is a model fit and the prediction for $x = 0$ for the partially classified subpopulations:

- with no disease
- with polyps
- with cancer


healthy
Variable
log(ratio)
log(ratio_h)
FITS_h
log(ratio)
x
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5


polyps
Variable
log(ratio)
log(ratio_p)
FITS_p
log(ratio)
x
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5


cancer
Variable
log(ratio)
log(ratio_
FITS_c
log(ratio)
x
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5

## prediction results for the Sydney screening study:

Table: *screening of 49927 subjects in Sydney for bowel cancer with partial verification of disease status; predictions using the log-linear model are in red; for comparison, Chao's estimator* $\hat{f}_0 = \frac{m-1}{m}\frac{f_1^2}{2f_2}$ *is given in brackets in blue*

| status | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| healthy | 15937(1990) | 1123 | 264 | 103 | 35 | 25 | 17 |
| polyps | 10638(1014) | 772 | 245 | 108 | 72 | 45 | 69 |
| cancer | 332(33) | 46 | 27 | 26 | 33 | 39 | 57 |
| total | 26907(3037) | 1941 | 536 | 237 | 140 | 109 | 143 |
| total | 46553 | 1941 | 536 | 237 | 140 | 109 | 143 |

## how well does the method work in practice?

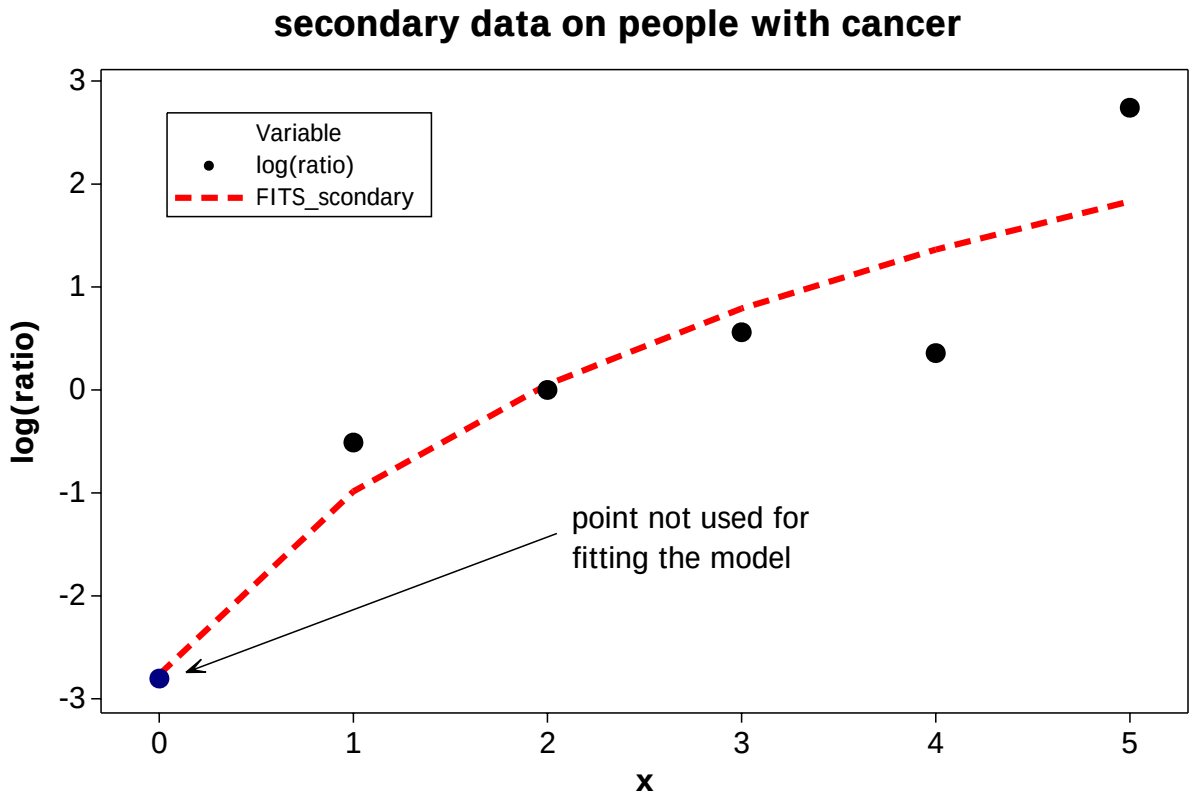
there are fully classified data available

- Lloyd and Frommer (2004) present a table with fully classified data
- a subset of 125 of the positive patients with confirmed cancer agreed to repeat the procedure 4 to 10 days after the primary test (secondary data)
- hence for all test results **including the test-negatives** the disease status is known

## results of for fully classified data

Table: *distribution of count of test-positives for a repeated diagnostic testing of 125 subjects with cancer*

| status | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| cancer | 25 | 8 | 12 | 16 | 21 | 12 | 31 |


secondary data on people with cancer
Variable
log(ratio)
FITS_scondary
point not used for
fitting the model
log(ratio)
x
3
2
1
0
-1
-2
-3
0
1
2
3
4
5

## results of for fully classified data

Table: *distribution of count of test-positives for a repeated diagnostic testing of 125 subjects with cancer (second row includes the predicted number of test-negatives using the model - in brackets Chao's estimator is given)*

| status | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| cancer | 25 | 8 | 12 | 16 | 21 | 12 | 31 |
| prediction | **21(2)** | 8 | 12 | 16 | 21 | 12 | 31 |

## identifiability?

easy to check

- let $\mathbf{Y} = (Y_0, Y_1, ..., Y_{m-1})^T$ and write model as
- $$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

  where $\mathbf{X}$ is the design matrix
- identifiability can be checked if the design matrix is of **full rank**
- hence, identifiability is reduced to the question of the identifiability of the regression model under consideration

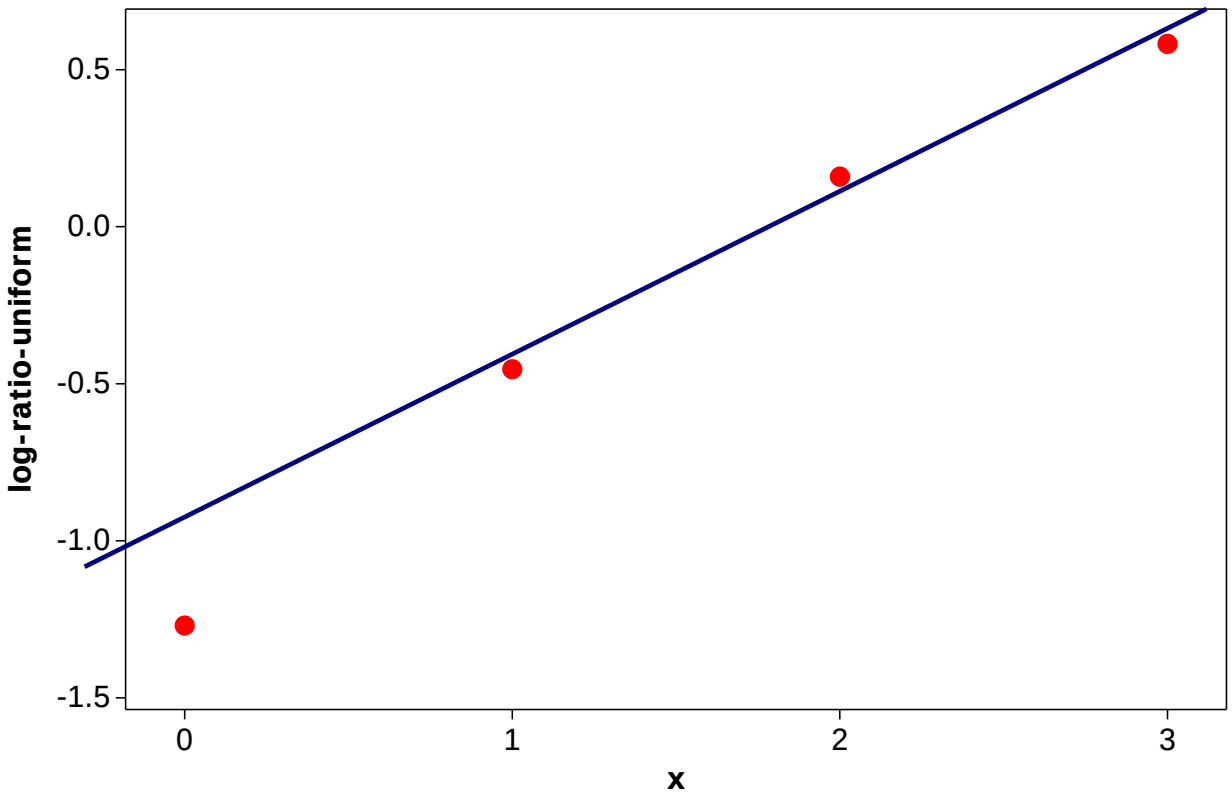
# example of Link (2003)

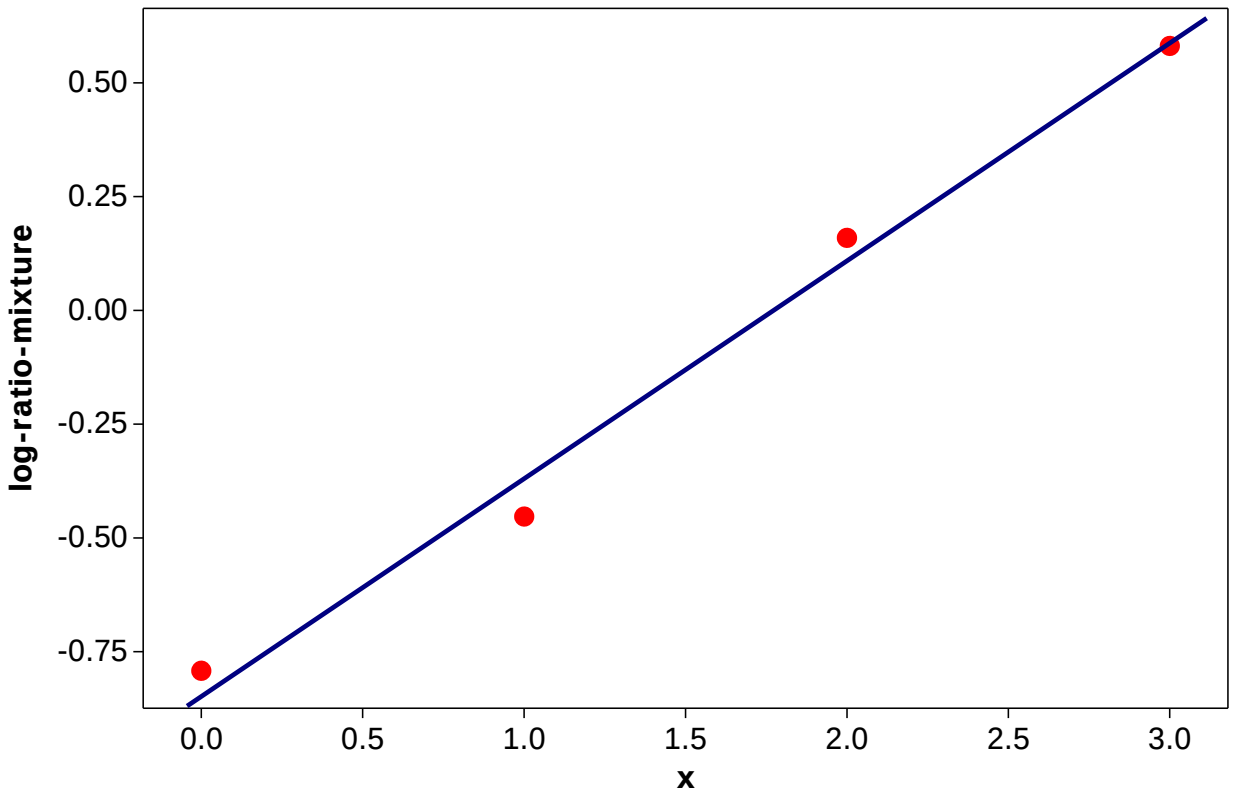
Table: *untruncated and truncated count distributions*

| model | probability | count $j$ 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| **uniform** | $p_j$ | 0.227 | 0.255 | 0.243 | 0.190 | 0.085 |
| | $p_j/(1-p_0)$ | - | 0.329 | 0.315 | 0.246 | 0.110 |
| **2 pt. mixture** | $p_j$ | 0.154 | 0.279 | 0.266 | 0.208 | 0.093 |
| | $p_j/(1-p_0)$ | - | 0.329 | 0.315 | 0.246 | 0.110 |

# regression approach can help to select suitable distributions

- need to restrict class of distributions under consideration
- regression approach can help to select the more plausible model
- as we see here ...


log-ratio-uniform
0.5
0.0
-0.5
-1.0
-1.5
0
1
2
3
x


log-ratio-mixture
0.50
0.25
0.00
-0.25
-0.50
-0.75
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x

## references (including all data used here)


- Lloyd, C.J. and Frommer (2004). Estimating the false negative fraction for a multiple screening test for bowel cancer when negatives are not verified. *Austr. N.Z.J. Stat.* **46**, 531-542.
- Lloyd, C.J. and Frommer (2004). Regression based estimation of the false negative fraction when multiple negatives are unverified. *J. Roy. Statist. Soc. Ser. C* **53**, 619-631.
- Lloyd, C.J. and Frommer (2008). An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification. *J. Roy. Statist. Soc. Ser. C* **57**, 89-102.

## related and recent papers


- Rocchetti, I., Bunge, J., Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics* **5**, 1512-1533.
- Böhning, D., Baksh, M.F., Lerdsuwansri, R., Gallagher, J. (2011). Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics* (in revision).
- Böhning, D. (2011). A regression approach to zero-truncated capture-recapture data. *Biometrika* (submitted).