School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

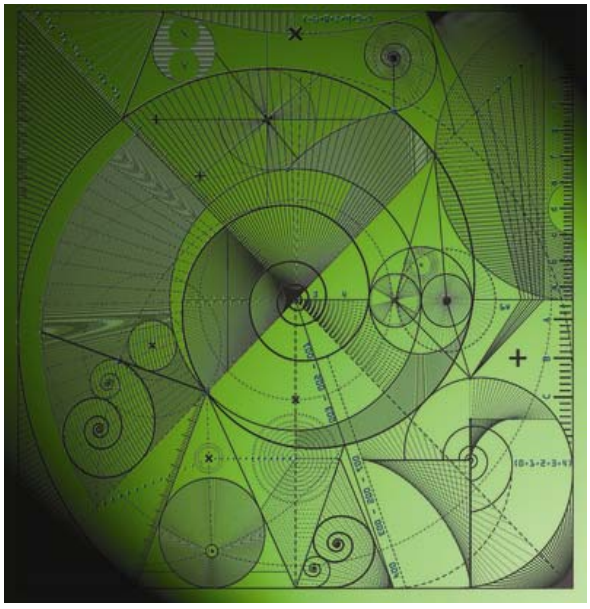
Preprint MPS-2011-04

25 March 2011

How many Cases Are Missed When Screening Human Populations for Disease?

by

Dankmar Böhning and Heinz Holling



How Many Cases Are Missed When Screening Human Populations for Disease?

Dankmar Böhning
Department of Mathematics and Statistics
School of Mathematical and Physical Sciences
University of Reading, Whiteknights
Reading, RG6 6BX, UK
email: d.a.w.bohning@reading.ac.uk

Heinz Holling
Statistics and Quantitative Methods
Faculty of Psychology and Sports Science
University of Münster, Münster, Germany
email: holling@psy.uni-muenster.de

March 25, 2011

Abstract

Human populations are frequently screened for specific diseases with the aim to detect the disease early when it is easier to treat and cure. However, only for those at risk (at risk being defined as having a positive screening test result) it is meaningful to verify the disease status. As a result a large portion of the screened population remains unverified in their disease status. We investigate techniques to estimate the amount of disease present in the unverified population. In particular, we consider a situation in which a specific screening test (here for the presence of bowel cancer) is applied several times and we focus on the count of times the test has been positive for each subject. A method is suggested to estimate the number of individuals with cancer but having a negative screening test result at all times. The suggested technique is based upon a simple log-linear model for the ratios of neighboring probabilities of the number of times the test has been positive. The technique is applied to publicly available data of a screening study on bowel cancer in Sydney and it is demonstrated that the method provides realistic estimates of the number of missed cancer cases and performs superior to other available techniques.

Some key words: capture-recapture, screening with partial verification of disease status, ratio estimator, zero–truncated model

1 The problem

Human populations are frequently screened for specific diseases with the aim to detect the disease early when it is easier to treat and cure. An example is screening for bowel cancer. Bowel cancer can develop without any early warning signs and can grow on the inside wall of the bowel for several years before spreading to other parts of the body. Often very small amounts of blood leak from these growths and pass into the bowel motion before any symptoms are noticed. A test called Faecal Occult Blood Test (FOBT) can detect these small amounts of blood in the bowel motion. The FOBT looks for blood in the bowel motion, but not for bowel cancer itself. Screening for bowel cancer using a FOBT is a simple non-invasive process that can be done in the privacy of your own home. No screening test is 100% accurate, in fact, a single application of the kit test might have low sensitivity. However, it is thought that a repeated replication of the diagnostic test over a number of days will help to identify most cases of cancer.

Over several years, from 1984 onwards about 50000 subjects were screened for bowel cancer at the St Vincent's Hospital in Sydney (Australia). The relevant references are Lloyd and Frommer (2004a, 2004b, 2008). The screening test comprises a sequence of 6 binary diagnostic tests which all are self-administered on 6 successive days. Each records the absence or presence of blood in faeces. If participants in the screening programme have their true disease status determined it is said the have been *verified*. In the case of this screening study it was done by physical examination, sigmoidoscopy and colonoscopy. Out of exactly 49927 participants, 46553 tested negatively on all six tests and were not further assessed with the implicit diagnosis that they were cancer-free. In other words, these 46553 participants remained unverified. Out of the other 3374 subjects who tested positively at least once, 3106 were examined and their true disease status determined with one of the following outcomes: healthy, polyps, cancer. 268 subjects who tested positively were lost to the study (Lloyd and Frommer 2008). The results are publicly available and presented here as Table 1.

Table 1: *Screening of 49927 subjects in Sydney for bowel cancer with partial verification of disease status (Lloyd and Frommer 2008)*

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
total	46553	1941	536	237	140	109	143

2 Conventional binomial model, diagnosing and modelling binomial heterogeneity

We are interested in determining the distribution of number of positive tests per subject. Assuming that there is a homogeneous probability θ that a single of the applied diagnostic tests is positive and that the 6 tests are applied independently, then number of positive tests X per subject follows the binomial distribution

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

where $n = 6$ is the number of tests per subject. Let us consider ratios

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

leading to

$$r_x = a_x\frac{p_{x+1}}{p_x} = \frac{x+1}{n-x}\frac{p_{x+1}}{p_x} = \frac{\theta}{1-\theta}, \tag{2}$$

where $a_x = (x+1)/(n-x)$, showing that the ratio r_x is constant with varying count x. It is straightforward to estimate $r_x = a_x\frac{p_{x+1}}{p_x}$ by

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

where f_x is the frequency of count x and $N = f_0 + f_1 + ... + f_n$. The graph $x \to \hat{r}_x = a_x\frac{f_{x+1}}{f_x}$ can be used as a *diagnostic device* for the binomial and is called the *ratio plot*. If the ratio plot shows a pattern of a horizontal line, it can be taken as indicative for the presence of a binomial distribution. This is demonstrated in Figure 1 for simulated data from a binomial with trial size parameter $n = 6$. The ratio plot shows clear evidence for a binomial distribution. If we apply the concept of the ratio plot to the Sydney screening data of Table 1, we see that there is no evidence of a horizontal line (see Figure 2). Instead, we observe in the ratio plot a monotone pattern arising. Indeed, if θ is distributed with arbitrary density $g(\theta)$, then

$$p_x = \int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}g(\theta)d\theta \tag{3}$$

has the monotonicity property (Böhning, Baksh, Lerdsruwansri, Gallagher 2011)

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

and, hence, the ratio plot is must be monotone increasing.

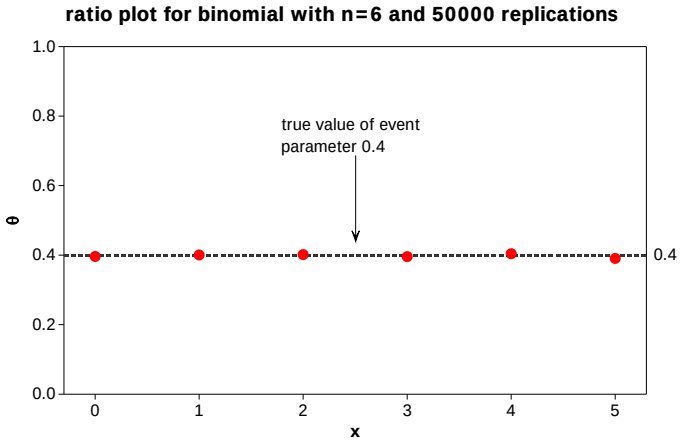


Figure 1: *Ratio plot for 50000 simulated binomial counts with event parameter* $\theta = 0.4$ *and trials size paramet* $n = 6$

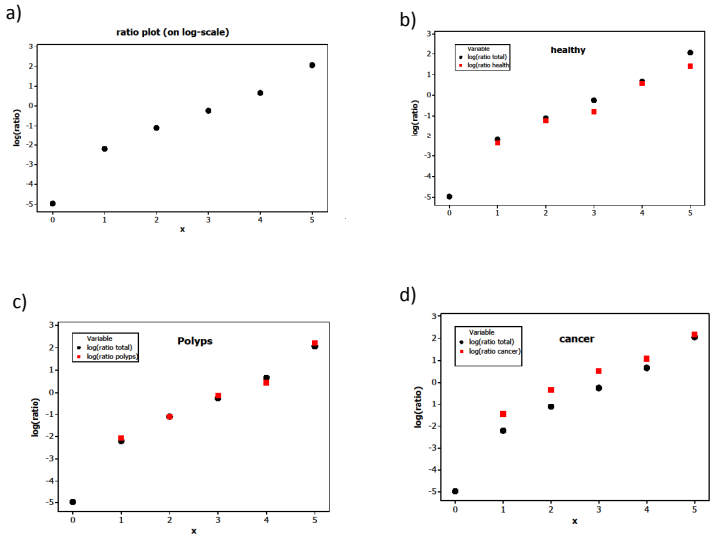


Figure 2: *Ratio plot for Sydney screening study: a) unclassified, b)healthy only (red squares), c) polyps only (red squares), d) cancer only (red squares)*

The form of the monotonicity in the pattern of the ratio plot in Figure 2 suggests to consider a more explicit modelling of $\log \hat{r}_x$ in its dependence from x. As it turns out, the model

$$\log \hat{r}_x = \alpha + \beta \log(x+1) + \epsilon_x \tag{4}$$

arises as a simple, quite reasonable model with only two parameters involved, namely α and β. We can find estimates of α and β using weighted regression estimates where the weights are found as the inverses of $1/f_{x+1} + 1/f_x$ (Böhning 2008). Figure 3 shows that the log-linear model (4) not only fits well for the unclassified data, but also for the partially classified data. Note that it is one of the advantages of the approach that the model can be fitted for the unclassified data as well as for the zero-truncated classified data (see Figure 3 a) – d).

To provide a more formal assessment of the model we consider the fitted values

$$\log \hat{r}'_x = \hat{\alpha} + \hat{\beta} \log(x+1),$$

for $x = 0, 1, ..., n-1$. Given $\hat{r}'_0, ..., \hat{r}'_{n-1}$, we need to determine $\hat{p}_0, ..., \hat{p}_n$. This can be accomplished as follows. By definition, the fitted probabilities have to satisfy the recursion $\hat{p}_{x+1} = (\hat{r}'_x/a_x)\hat{p}_x$ for $x = 0, ..., n-1$ as well as $\sum_{x=0}^{n} \hat{p}_x = 1$, so that

$$1 = \hat{p}_0 + ... + \hat{p}_n = p_0[1 + \hat{r}'_0/a_0 + (\hat{r}'_0/a_0)(\hat{r}'_1/a_1) + ... + \prod_{x=1}^{n-1} \hat{r}'_x/a_x],$$

which leads to

$$\hat{p}_0 = \frac{1}{1 + \sum_{i=0}^{n-1} \prod_{x=0}^{i} \hat{r}'_x/a_x}. \tag{5}$$

Table 2: *Observed and fitted frequencies of the unclassified Sydney screening data with various models fitted: log-linear (4), binomial model (1), and beta-binomial*

model	0	1	2	3	4	5	6	χ^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	398.93
log-linear	46418.7	1968.5	443.1	235.7	202.8	211.2	178.9	96.37
observed	46553	1941	536	237	140	109	143	-

The remaining fitted probabilities can be determined using the recursion $\hat{p}_{x+1} = (\hat{r}'_x/a_x)\hat{p}_x$ for $x = 0, ..., n-1$. The fitted frequencies for the unclassified Sydney screening data, $\hat{f}_x = \hat{p}_x N$ are found in row 4 of Table 2. The fit is good with a $\chi^2 = \sum_{x=0}^{n} \frac{(f_x - \hat{f}_x)^2}{\hat{f}_x} = 96.37$ by $df = 7 - 1 - 2 = 4$ degrees of freedom. For comparison, we have also included the fitted frequencies under a binomial model (1) which is evidently entirely unsatisfactory. An improved fit can be found by applying the *beta-binomial* model with $g(\theta)$ in (3) provided as the beta-density

$$g(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where α and β are unknown parameters and $\Gamma(t) = \int_0^\infty y^{t-1}\exp(-y)dy$ is the Gamma function. Note that the beta-binomial is frequently used since the marginal can easily be worked out to be

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

The fit of the beta-binomial with $\chi^2 = 398.93$ is evidently a lot better than the fit for the binomial, but clearly inferior to the fit of $\chi^2 = 96.37$ provided by the log-linear model. Hence, we will focus in the following on the log-linear model. The panels b), c) and d) in Figure 3 show that the log-linear model is also suitable for the partially classified populations with cancer, with polyps, and those that are healthy.

3 Predicting the number of cases missed

We are now considering fitting the log-linear model (4) to the partially classified data as

- being healthy
- having polyps
- having cancer.

For each of the three populations we fit the log-linear model

$$\log \hat{r}_x = \alpha + \beta \log(x+1) + \epsilon_x$$

for $x = 1, ..., n-1$. Note that model is suitable for both situations, untruncated and zero-truncated count data – as we have here. Clearly, we are interested in the *predicted value* $\hat{r}'_0 = \exp(\hat{\alpha})$ which motivates the estimator

$$\hat{f}_0 = a_0 f_1 \exp(-\hat{\alpha}), \tag{6}$$

since $\hat{r}'_0 = a_0\hat{f}_1/\hat{f}_0$, and replacing $\hat{f}_1$ by the observed frequency f_1 provided the estimate in (6). The specific estimates for the three populations is given in Table 3. As a standard comparative estimator we provide Chao's lower bound estimate which is achieved as

$$\hat{f}_{0,C} = \frac{n-1}{n}\frac{f_1^2}{2f_2},$$

which provides an asymptotic lower bound for $E(f_0)$. This lower bound is given in brackets in column 2 of table 3. Typically, Chao's lower bound provides only 10% of the size of the log-linear estimator. Still, also our estimates are lower bounds since 15937+10638+332=26907 do not match the marginal frequency 46553. The shape of the graphs in Figure 3 suggest that we likely underestimate the frequency of zero-counts in the healthy population and in the population with polyps, not so much the in the population with cancer.

The estimate for the frequency of missed cancer cases seems realistic. Values for the sensitivity of an individual application of a FOBT ranges between $q = 0.1 - 0.3$. This leads to a probability for the diagnostic procedure finding at least one positive result in n successive applications of $1-(1-q)^n$. This would mean to expect the frequency of missed subjects with cancer to range between $258 - -487$. We estimate it with 332 which is well in this range.

Table 3: *screening of 49927 subjects in Sydney for bowel cancer with partial verification of disease status: predicted frequencies of zero-count test results for each of the three partially classified populations are in bold (second column)*

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	46553	1941	536	237	140	109	143

4 Prediction for a subset of fully classified data

Lloyd and Frommer (2004) present also a table of cancer patients with a repeated application of the diagnostic procedure. In fact, a subset of 125 of the *positive patients with confirmed cancer* agreed to repeat the procedure 4 to 10 days after the primary test which we call the *secondary data*. Hence for all test results *including the test-negatives* the disease status is known (Table 4). It might be

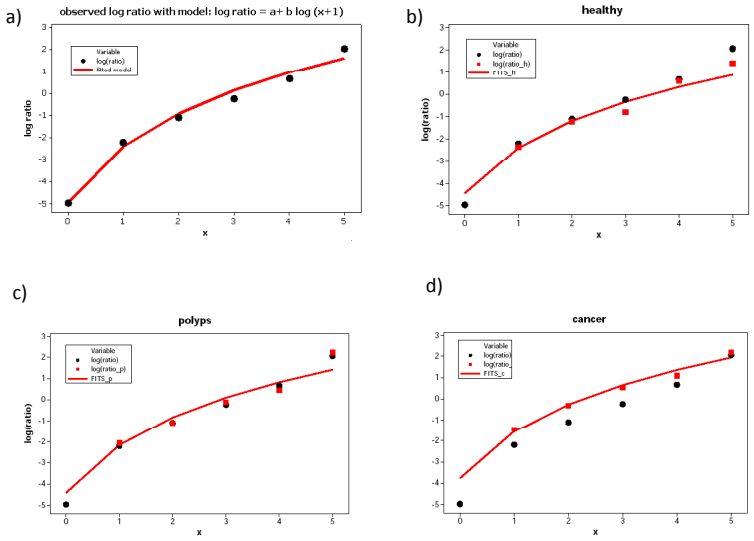


Figure 3: *Ratio plot with fitted log-linear model for Sydney screening study: a) unclassified with fitted line, b)healthy only (red squares) with fitted model (solid red line) , c) polyps only (red squares) with fitted model (solid red line) , d) cancer only (red squares) with fitted model (solid red line); for b)-d) the black dots indicate the unclassified data*

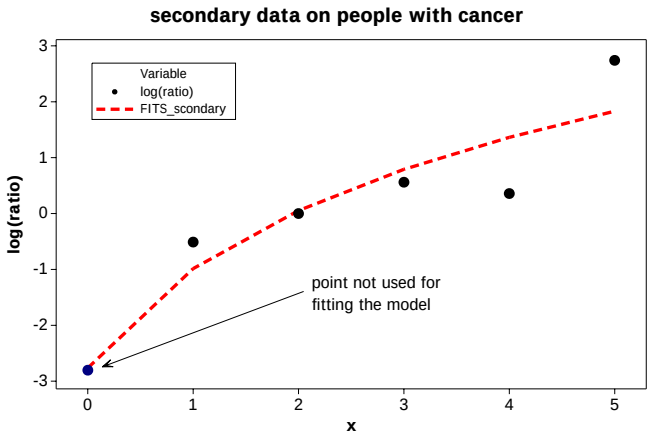


Figure 4: *Ratio plot with fitted log-linear model for secondary Sydney screening data with confirmed cancer*

interesting to see how the log-linear estiamte (**??**) performs when the known frequency f_0 is ignored. A model fit ignoring f_0 is presented in Figure 4 which appears to be reasonable. The corresponding estimate from (6) is $\hat{f}_0 = 21$ which compares favorably with the observed value of $f_0 = 25$. Chao's estimate is poor with $\hat{f}_{0,C} = 2$.

Table 4: *Distribution of count of test-positives for a repeated diagnostic testing of 125 subjects with cancer*

	number of positive tests						
x	0	1	2	3	4	5	6
f_x	25	8	12	16	21	12	31

References

[1] Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statistical Methodology* **5**, 410–423.

[2] Bunge, J. and Fitzpatrick, M. (1993). Estimating the Number of Species: a Review. *Journal of the American Statistical Association* **88**, 364–373.

[3] Chao, A. (1987). Estimating the Population Size for Capture-Recapture data with Unequal Catchability. *Biometrics* **43**, 783–791.

[4] Dorazio, R.M. and Royle, J.A. (2005). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351–364.

[5] 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.

[6] 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.

[7] 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.