A FIXED-INVERSE BINARY MISCLASSIFICATION MODEL UNDER POSSIBLE FALSE-POSITIVE MISCLASSIFICATION

Asmerom Tesfamichael, Advisor: Dr. Kent Riggs

Stephen F Austin State University – Department of Mathematics and Statistics

Abstract

- In this project, we develop a particular statistical model for binary data that allows for the possibility of false-positive misclassification. To account for the misclassification, the model incorporates a two-stage sampling scheme.
- Next, we apply maximum likelihood methods to find estimators of the primary prevalence parameter p as well as the false-positive misclassification rate parameter ϕ. In addition, we derive confidence intervals for p based on inverting Wald, score and likelihood ratio statistics.
- Also, we graphically compare coverage and width properties of the Wald-based, score-based, and likelihood ratio-based confidence intervals for p through a Monte Carlo simulation. The simulation study is done under different parameter and sample size configurations. Also, we apply the newly-derived confidence intervals for p to a real data set.

Introduction

- Due to practical reasons such as cost and time savings, fallible classifiers which are prone to error are used to classify binary data.
- Misclassification may result in false-negatives or false-positives.
- Misclassification errors may distort results of statistical analysis.
- Models that account for misclassification have been developed to compensate for the effect of errors.
- The misclassification rate parameter is a measurable feature of a statistical model that accounts for misclassification.
- Different applications requires different statistical models, which have specific advantages and limitations.
- Better estimation can be done by an infallible device, but at a higher cost.
- A double sampling scheme using both fallible and infallible devices may be used at a reasonable cost , while properly accounting for misclassification.
- We consider a model that allows only for false-positive misclassification, which treats the first sample of a two-stage sampling scheme as fixed and the second stage of the scheme as random (inverse sampling).

Two-Stage Sampling Scheme

The double sampling scheme involves the use of a fallible (cheap) and infallible (expensive) classifier in two stages in an effort to appropriately estimate p and the rate(s) of misclassification. The first stage involves the use of a fallible classier that is prone to producing false-positives under fixed sampling. The second stage involves using an infallible and fallible classifier under inverse sampling technique. For insight into this two-stage scheme consider the following example (* denotes false-positive):

Stage	Pop.	0	1	1	0	0	0	0	0	1	1	0	1	0	1	0	1
First	Fallible	0	1	1	1*	1*	0	0	1								
Sec.	Fallible									1	1	1*	1	0	1	1*	1
	Infallible									1	1	0	1	0	1	0	1

Fixed-Inverse Binary Misclassification Model

For the two stage sampling scheme, define the following counts:

y = # of observations labeled success after m trials of the fallible device in stage 1,

n_{00} = # of observations labeled failure by both fallible and infallible devices in stage 2,

n_{10} = # of observations labeled "success" by fallible device but "failure" infallible device in stage 2,

n_{11} = # of observations labeled success by both fallible and infallible devices in stage 2.

For example above, $y = 5$, $n_{00} = 1$, $n_{10} = 2$, and $n_{11} = 5$.

- **Distributional Assumptions**

The Binomial distribution and Negative Multinomial distribution are used to model the counts y, n_{00}, n_{10} :

$$f_1(y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}$$

and

$$f_2(n_{00}, n_{10}) = \frac{(n_{11}+n_{00}+n_{10}-1)!}{(n_{11}-1)!\,n_{00}!\,n_{10}!}((1-p)(1-\phi))^{n_{00}}(\phi(1-p))^{n_{10}}p^{n_{11}}$$

where,

$n = n_{00} + n_{10} + n_{11}$ is the random sample size needed to observe n_{11} successes labeled by both fallible and infallible classifiers in stage 2,

p = probability infallible device yield success,

ϕ = probability fallible device yield false-negative,

π = probability fallible device labels an observation as success.

- **Maximum Likelihood Estimators**

$$\hat{p} = \frac{n_{11}(n_{11}+n_{10}+y)}{(n_{11}+n_{10})(n_{00}+n_{10}+n_{11}+m)}$$

and

$$\hat{\phi} = \frac{n_{10}(n_{10}+n_{11}+y)}{(n_{11}+n_{10})(n_{00}+n_{10}+n_{11}+m)(1-\hat{p})}$$

- **Large Sample Confidence Intervals for p**

  - **Wald CI**: $\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{I^{11}(\hat{p},\hat{\phi})}$
  - **Score CI:** values of p that satisfy :
  $$[u_p(\hat{\phi}_p)]^2\, I^{11}(p,\hat{\phi}_p) \le \chi^2{}_1(\alpha)$$
  - **Likelihood CI:** values of p that satisfy :
  $$2\left(l(\hat{p},\hat{\phi}) - l(p,\hat{\phi}_p)\right) \le \chi^2{}_1(\alpha)$$

where,

- $Z_{\frac{\alpha}{2}}$ is $(1-{}^{\alpha}/_{2})$ percentile for the standard normal distribution
- $u_p(\hat{\phi}_p) = \frac{\partial l}{\partial p}$ at $\hat{\phi}_p$
- $\chi^2{}_1(\alpha)$ is $(1-\alpha)$ percentile of a chi-squared distribution with one degree of freedom
- $I^{11}(p,\hat{\phi}_p)$ is the $(1,1)$ element of the inverse of Fisher's Information Matrix
- l is the log likelihood function

Motivating Example

- The western blot procedure (WBP) is one diagnostic test of the herpes simplex virus. Out of 693 women tested, the WBP yielded that 375 had the virus. (Hildeshiem and Boese)
- Under the binomial model (not accounting for misclassification), the maximum likelihood estimator (MLE) and Wald confidence interval for p is

Parameter	Estimate	Wald C.I
p	0.541	(0.504,0.578)

It turns out that WBP is a fallible detector of the virus, hence the estimate above is biased.

- Another procedure called the Refined Western Blot Procedure (RWBP) is accurate and we will consider it as an infallible classifying device.
- Also, from Hildesheim's data we will only consider the false positives from stage 2 and allow the false negatives to be absorbed into n_{11}.
- From the two stages, the following counts were observed:
  $y = 375$, $m = 693$, $n_{00} = 13$, $n_{10} = 3$, and $n_{11} = 23$.
- Under the Fixed-Inverse Binary Misclassification Model, the MLE's for p and ϕ and confidence intervals for p are

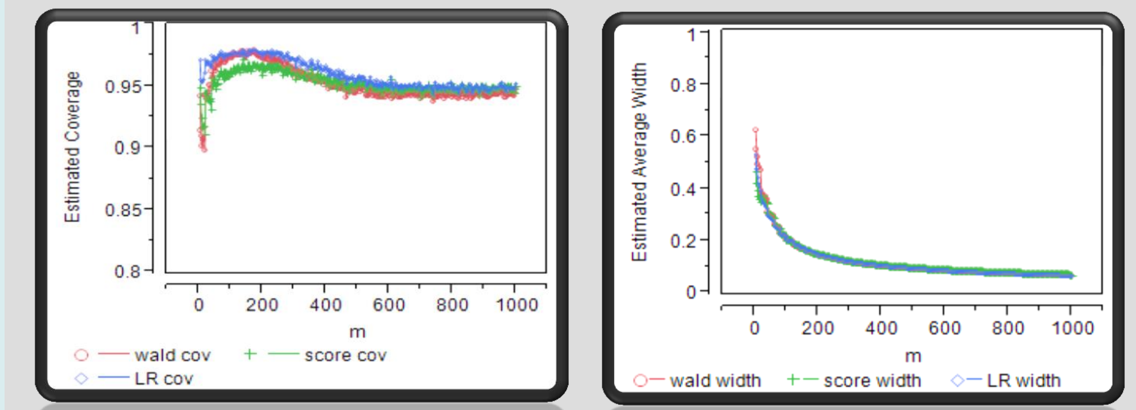
Parameter	Estimates	Wald C.I *	Score C.I.*	Lik.-Ratio C.I *
p	0.485	(0.4102,0.5589)	(0.3899,0.5793)	(0.4279,0.5413)
ϕ	0.123	Not yet	Not yet	Not yet

- The estimate of p under Fixed-Inverse Binary Misclassification Model is smaller than under the binomial model, which is intuitive because the misclassification rate (ϕ) is around 12% . Hence, the estimate of p under the binomial model is likely overestimated due to false-positives generated by using only the fallible classifier.

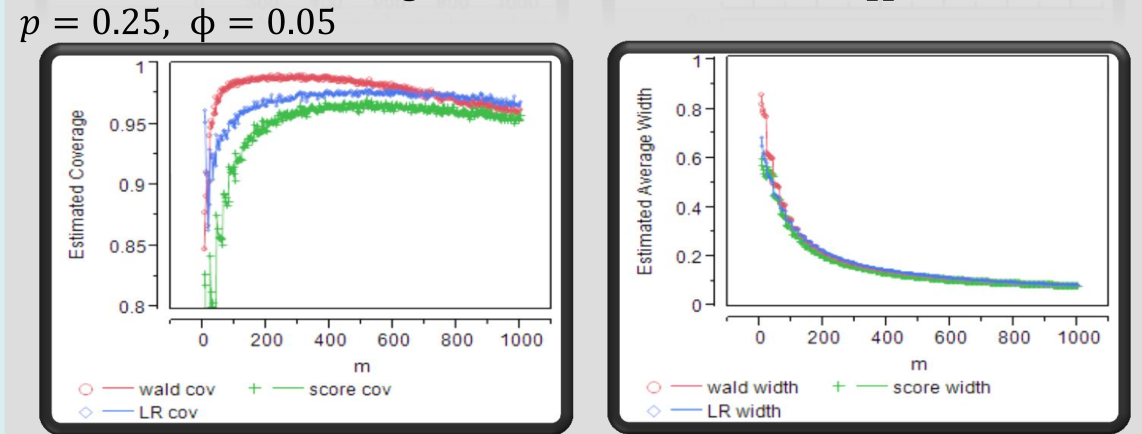
Simulation Results

- Next, we consider the coverage and width properties of the three confidence intervals when estimating p.
- Here we consider two configurations of ratios of the fallible data to the infallible data: $n_{11} = 0.05m$ and $n_{11} = 0.4m$, while varying the parameters p and ϕ.
- All simulations were performed in SAS IML with a simulation size of 10,000 iterations.
- The nominal confidence level was 95%.

Simulation Results When $n_{11} = 0.05m$

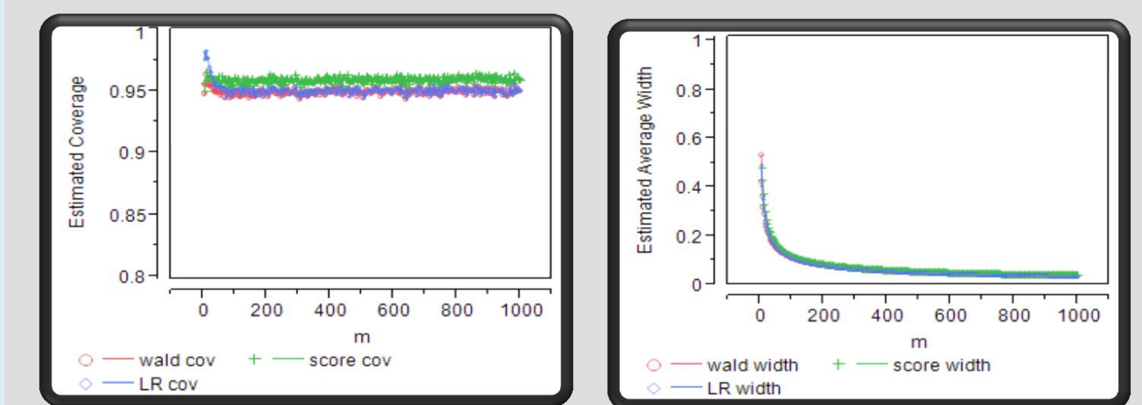


Estimated Actual Coverages and Actual Widths when $n_{11} = 0.05m$ and $p = 0.25,\ \phi = 0.05$

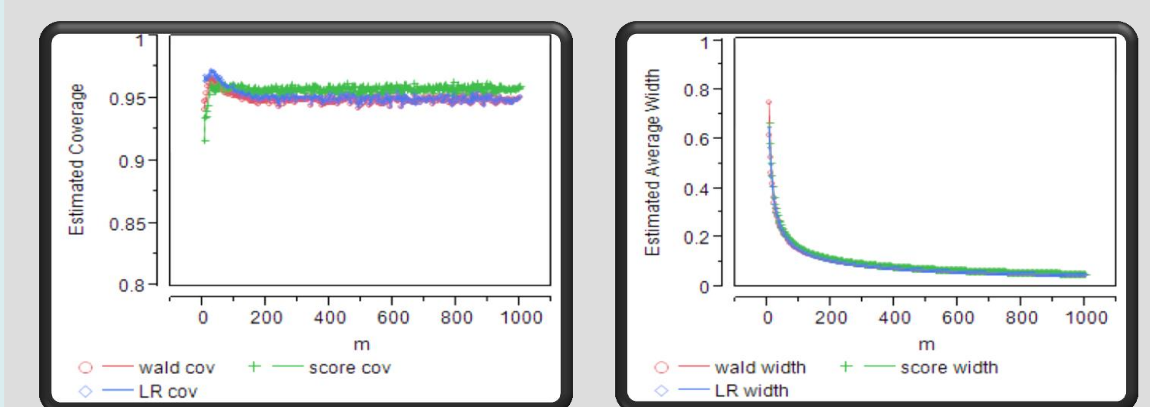


Estimated Actual Coverages and Actual Widths when $n_{11} = 0.05m$ and $p = 0.5,\ \phi = 0.05$

Simulation Results When $n_{11} = 0.4m$



Estimated Actual Coverages and Actual Widths when $n_{11} = 0.4m$ and $p = 0.25,\ \phi = 0.05$



Estimated Actual Coverages and Actual Widths when $n_{11} = 0.4m$ and $p = 0.5,\ \phi = 0.05$

Bibliography

1, Hogg V.R. Mckean W.J. and Craig, T.A., Introduction to Mathematical Statistics, sixth edition ed., Pearson Prentice Hall, 2005.
2, Boese, D., Young, D., and Stamey, J., (2006), "Confidence Intervals for a Binomial Parameter Based on Binary Data Subject to False- Positive Misclassification," Computational Statistics and Data Analysis, 50, 3369- 3385.
3, Tenennbein, A., (1970), "A Double Sampling Scheme for Estimating From Binomial Data with Misclassifications," Journal of American Statistical Association, 65 (331), 1350- 1361.
4, Hildesheim, A., Mann, V., Brinton, L.A., Szklo, M., Reeves, W.C., Rawls, W.E., (1991), "Herpes Simplex Virus Type 2: A Possible Interaction with Human Papillomavirus 16/18 in the Development of Invasive Cervical Cancer," International Journal of Cancer, 49, 335-340.