Journal of Data Science 13(2015), 623-636

Confidence Intervals for a Proportion Using Inverse Sampling when the Data is Subject to False-positive Misclassification

Kent Riggs [1]
[1] *Department of Mathematics and Statistics, Stephen F. Austin State University*

Abstract: Of interest in this paper is the development of a model that uses inverse sampling of binary data that is subject to false-positive misclassification in an effort to estimate a proportion. From this model, both the proportion of success and false-positive misclassification rate may be estimated. Also, three first-order likelihood-based confidence intervals for the proportion of success are mathematically derived and studied via a Monte Carlo simulation. The simulation results indicate that the score and likelihood ratio intervals are generally preferable over the Wald interval. Lastly, the model is applied to a medical data set.

Key words: Misclassification; Double sampling; Inverse sampling; Interval estimation; Likelihood methods

1. Introduction

Statistical inference of a proportion using binary data that is subject to misclassification has been a part of the statistical literature for some time. Bross (1954) demonstrated that traditional estimation of a proportion in the presence of misclassification produces biased estimators. Tenennbein (1970) accounts for the misclassification with a double sampling scheme. Lie et al. (1994) and Moors et al. (2000) considered the case where only false-negative counts are obtained. Boese et al. (2006) developed several interval estimators in the case when only false-positive counts are obtained.

All of the above authors used a fixed sampling scheme(s) in order to estimate the proportion of interest. While this approach often works well, it has been demonstrated by Tian et al. (2009) that fixed samples can sometimes be too small to effectively estimate the proportion of success when it is small. As a remedy to this problem, inverse sampling and the negative binomial distribution can be employed.

In this paper, I wish to develop a statistical model that uses a two-stage sampling scheme to estimate a proportion of success where the first stage is generated under inverse sampling and is subject to false-positives, while the second stage is generated under fixed sampling. The paper is organized as follows. In Section 2, the statistical model is developed, and in Section 3 maximum likelihood estimators (MLE) are presented for the proportion of success and false-positive rate. Interval estimators are also derived by inverting the Wald, score, and likelihood ratio statistics in Section 3. In Section 4, the coverage properties and average widths of the three interval estimators are compared using a Monte Carlo simulation. Finally, in Section 5, the three confidence intervals are applied to a real-world data example.

2. The False-Positive Model Using Inverse Sampling

The first stage in the two-stage sampling scheme involves the use of a fallible classifier that is prone to producing false-positives under inverse sampling. The second stage involves using an infallible classifier under fixed sampling. In conjunction with the fallible classifier, one can obtain the actual number of successes, actual number of failures, and actual number of false positives for this fixed sample in the second stage. The two stages are considered independent. The two-stage sampling scheme is very useful to appropriately estimate the main proportion of interest as well any false-positive misclassification parameter. Appropriate estimation of these parameters gives the researcher better insight into the attribute of interest and the rate at which a fallible classifier may produce a false-positive.

Let *Y* be the number of failures labeled by the fallible device in stage 1 until the *k*th "success" is observed. Hence, $Y \sim NegBin(k,\pi)$, where π is the probability the fallible device *labels* an observation as a "success". Once stage 1 is completed, the fixed sampling for stage 2 is implemented, where both the fallible and infallible devices make classifications on n observations. Let n_{00} be the number of observations labeled "failure" by both the fallible and infallible devices, n_{10} be the number of observations labeled "success" by the fallible device but labeled "failure" by the infallible device, and n_{11} be the number of observations labeled "success" by both the fallible and infallible devices. Thus, $n = n_{00} + n_{10} + n_{11}$, and

$$(n_{00}, n_{10}, n_{11}) \sim trinomial\left(n, (1-\phi)(1-p), \phi(1-p), p\right),$$

where ϕ is the probability of the fallible device yielding a false-positive, and p is the interested probability of success. Therefore, it can be shown that $\pi = p + (1-p)\phi$.

Table 1 provides an example of the two-stage sampling plan. Observations for both stages are gathered on different portions of the population. "0" denotes a failure, "1" denotes a success, and "1*" denotes a false-positive by the fallible classifier. For stage 1, $k = 5 \text{ and } y = 3$, while for stage 2, $n_{00} = 1, n_{10} = 2, \text{ and } n_{11} = 4$.

Table 1
Two-stage Sampling Scheme Example

	Population	0	1	1	1	0	1	0	0	1	1	0	1	0	1	0	…	1
Stage 1	Fallible	0	1	1	1	1*	1	0	0									
Stage 2	Fallible									1	1	1*	1	0	1	1*		
	Infallible									1	1	0	1	0	1	0		

Table 2 gives insight into the resulting probabilities from the different stage and classifying device combinations.

Table 2
Probabilities for Each Stage/Classifier Scenario

Stage	Infallible Device	Fallible Device	
		0	1
First	N/A	$1-\pi$	π
Second	0	$(1-\phi)(1-p)$	$\phi(1-p)$
	1	N/A	p

3. Estimation

Using samples from both stages results in the following likelihood function:

$$L(p,\phi) = trinomial\left(n,(1-\phi)(1-p),\phi(1-p),p\right)\times NegBin(k,\pi)$$

$$= \frac{n!}{n_{00}!n_{10}!n_{11}!}\left[(1-\phi)(1-p)\right]^{n_{00}}\left[\phi(1-p)\right]^{n_{01}}p^{n_{11}}\begin{pmatrix} y+k-1 \\ k-1 \end{pmatrix}\pi^{k}(1-\pi)^{y}$$

$$= \frac{n\,!}{n_{00}\,!n_{10}\,!n_{11}\,!}\Big[(1-\phi)(1-p)\Big]^{n_{00}}\Big[\phi(1-p)\Big]^{n_{01}} p^{n_{11}}\begin{pmatrix} y+k-1 \\ k-1 \end{pmatrix}$$
$$\times\Big[p+(1-p)\phi\Big]^{k}\Big[(1-\phi)(1-p)\Big]^{y}. \tag{1}$$

From (1), we can straightforwardly find Fisher's Expected Information Matrix, which is

$$I(p,\phi)=\begin{bmatrix} I_{pp} & I_{p\phi} \\ I_{\phi p} & I_{\phi\phi} \end{bmatrix} = -E\begin{bmatrix} \dfrac{\partial^2 \ell}{\partial p^2} & \dfrac{\partial^2 \ell}{\partial p \partial \phi} \\ \dfrac{\partial^2 \ell}{\partial \phi \partial p} & \dfrac{\partial^2 \ell}{\partial \phi^2} \end{bmatrix},$$

where $\ell(p,\phi)=ln\big(L\big)$ is the log-likelihood function. The elements of this matrix are given in the appendix. Using calculus, it can be shown that the MLE's for p and ϕ are $\hat{p}=\left(\dfrac{n_{11}}{n_{10}+n_{11}}\right)\left(\dfrac{k+n_{10}+n_{11}}{n+k+y}\right)$ and $\hat{\phi}=\left(\dfrac{n_{10}}{n_{10}+n_{11}}\right)\left(\dfrac{k+n_{10}+n_{11}}{(k+y+n)(1-\hat{p})}\right)$. Also, it can be shown that the restricted MLE of ϕ for a fixed p is given by

$$\hat{\phi}_p = \frac{-n_{10}+k(p-1)+n_{00}p+2n_{10}p+py}{2(p-1)(k+n_{00}+n_{10}+y)}$$
$$-\frac{\sqrt{k^2(p-1)^2+2k(p-1)(-n_{10}+p(n_{00}+y))+(n_{10}+p(n_{00}+y))^2}}{2(p-1)(k+n_{00}+n_{10}+y)}.$$

Next, we derive confidence intervals for p by inverting the appropriate Wald, score, and log-likelihood statistics. Confidence intervals for ϕ can be derived in a similar manner, but we will not pursue them in this paper. All three interval estimators are based on first order asymptotic approximations.

The Wald-based interval estimator for p uses the unrestricted MLE's $\hat{p}$ and $\hat{\phi}$. For sufficiently large samples, $\hat{p}$ □ $N\big(p, I^{11}(p,\hat{\phi})\big)$, where $I^{11}(\hat{p},\hat{\phi})$ is the (1,1) element of $\big(I(\hat{p},\hat{\phi})\big)^{-1}$. Hence, an approximate $100(1-\alpha)\%$ confidence interval is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{I^{11}(p,\hat{\phi})}, \tag{2}$$

where $z_{\alpha/2}$ is the 100(1-$\alpha/2$) percentile for the standard normal distribution.

A score-based confidence interval for p involves inverting the score statistic, which requires use of $\hat{\phi}_p$. For sufficiently large samples, $\left[u_p(\hat{\phi}_p)\right]^2 I^{11}(p,\phi_p) \sim \chi_1^2$,

where $u_p(\hat{\phi}_p) \equiv \left.\frac{\partial \ell}{\partial p}\right|_{\hat{\phi}_p}$ and $I^{11}(p,\hat{\phi}_p)$ is the (1,1) element of $\left(I(p,\hat{\phi}_p)\right)^{-1}$. Therefore, an approximate $100(1-\alpha)\%$ confidence interval is composed of the values of p that satisfy

$$\left[u_p(\hat{\phi}_p)\right]^2 I^{11}(p,\phi_p) \le \chi_1^2(\alpha), \tag{3}$$

where $\chi_1^2(\alpha)$ is the $100(1-\alpha)$ percentile of a chi-square distribution with one degree of freedom. The expression on the left-hand side of (3) is complicated, and solutions to (3) must be found numerically.

The likelihood ratio confidence interval for p involves inverting the log-likelihood statistic. Note that $2\left(\ell(\hat{p},\hat{\phi}) - \ell(p,\phi_p)\right) \sim \chi_1^2$, for sufficiently large samples. Hence, an approximate $100(1-\alpha)\%$ confidence interval is composed of the values of p that satisfy

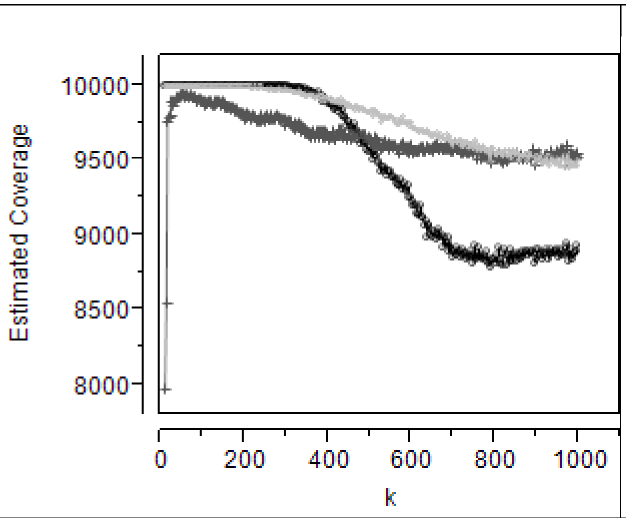
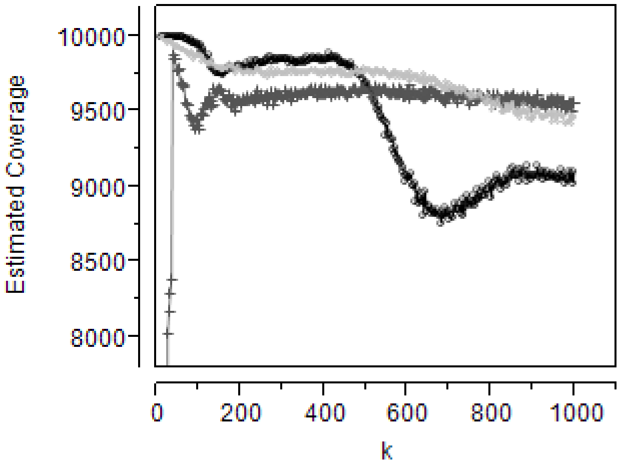
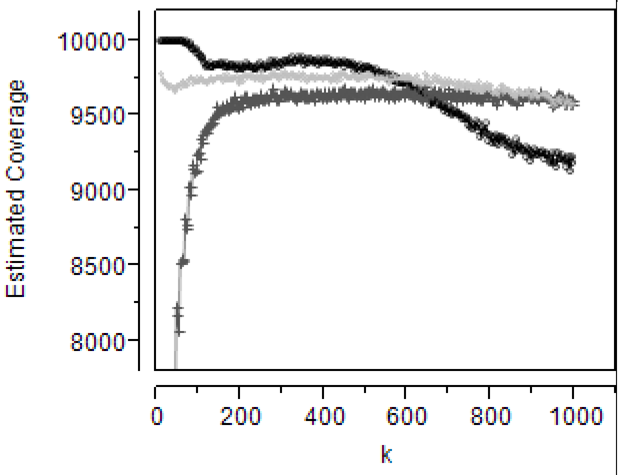
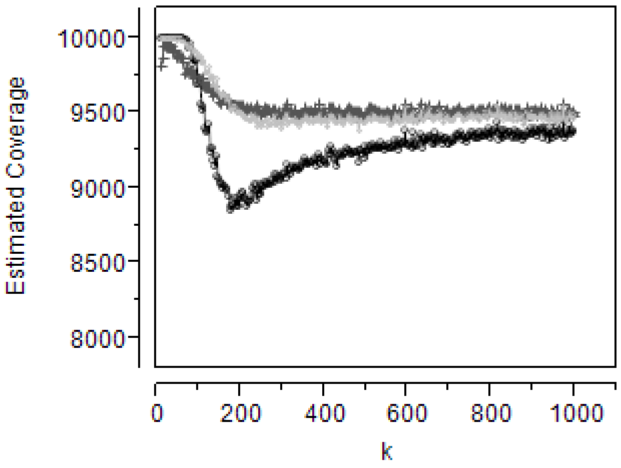
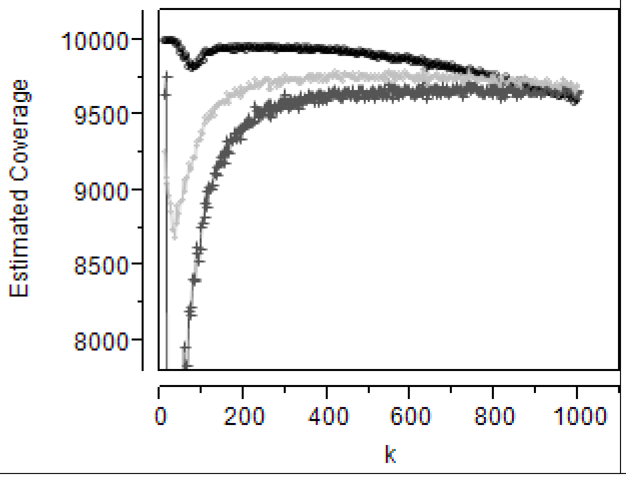
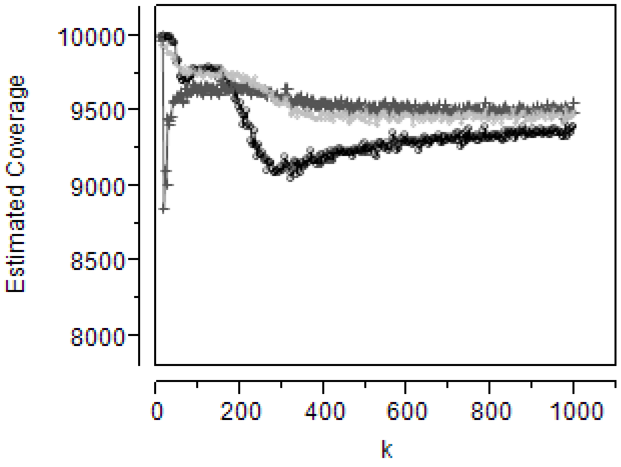
$$2\left(\ell(\hat{p},\hat{\phi}) - \ell(p,\phi_p)\right) \le \chi_1^2(\alpha). \tag{4}$$

As in (3), the values of p that satisfy (4) must be determined numerically.

4. Simulation Study

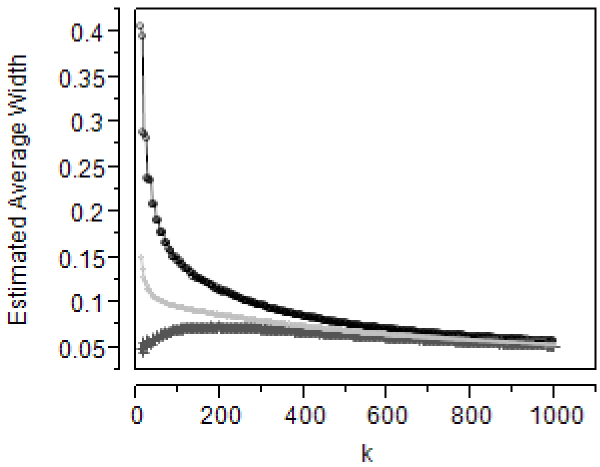
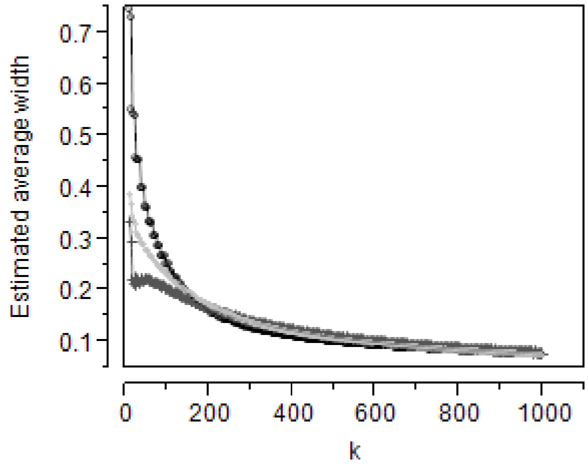
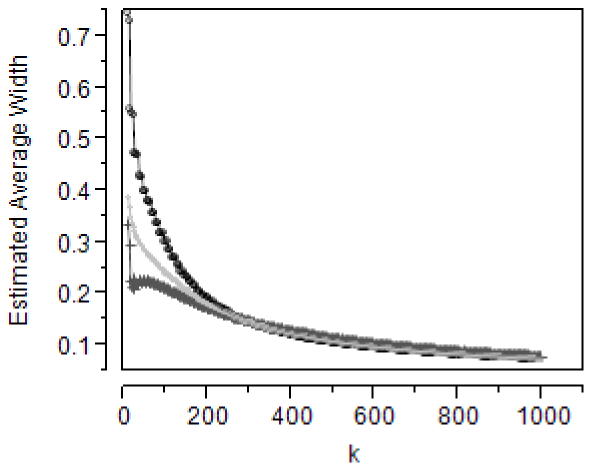
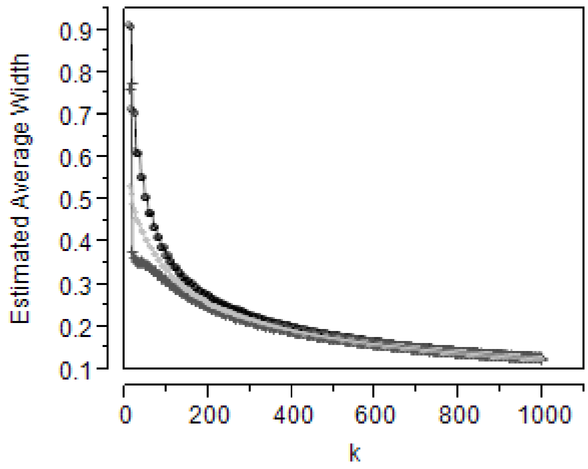
Now we consider coverage and width properties of the interval estimators described in (2) - (4). First, we examine the coverage and width properties when $n = 0.1k$ for various combinations of p and ϕ. All simulations were performed in SAS IML V9.3 with 10,000 iterations for each parameter and sample size configuration. Figures 1 and 2 give plots that summarize the coverage and average width properties of confidence intervals (2) - (4). The nominal confidence level is 95% and all estimated coverages have standard errors less than 0.005. The maximum standard error for estimated average widths in Figure 2 is 0.00231.

	$\phi = 0.05$	$\phi = 0.25$

$p = 0.05$		
$p = 0.25$		
$p =0.50$		

○ — Wald_Coverage + — Score_Coverage ◇ — Likelihood_Ratio_Coverage

Figure 1:Coverage Plots for the Wald, Score, and Likelihood Ratio Confidence Intervals when n = 0.1k

	$\phi = 0.05$	$\phi = 0.25$
$p = 0.05$		
$p = 0.25$		

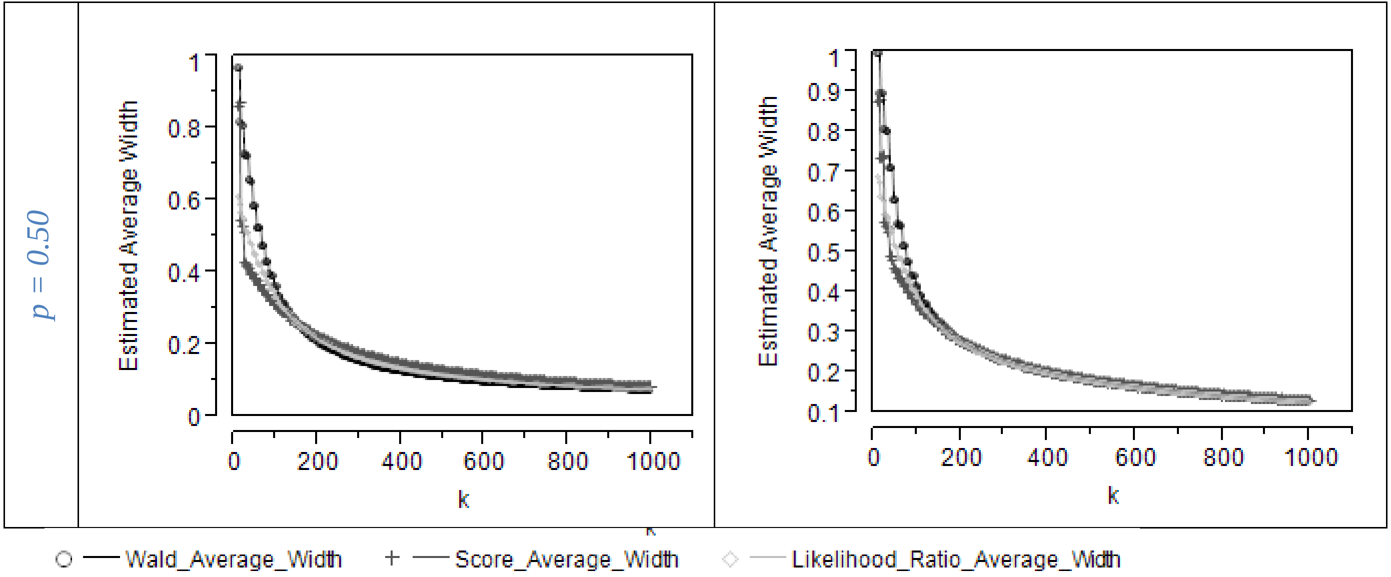
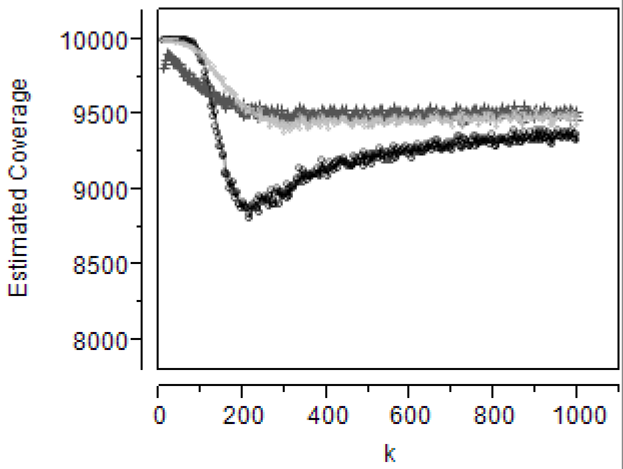
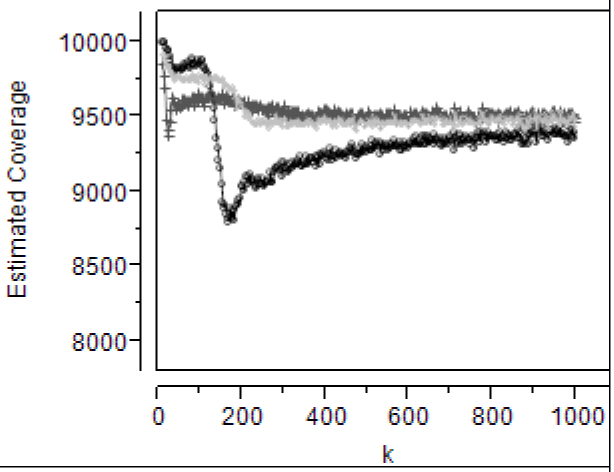
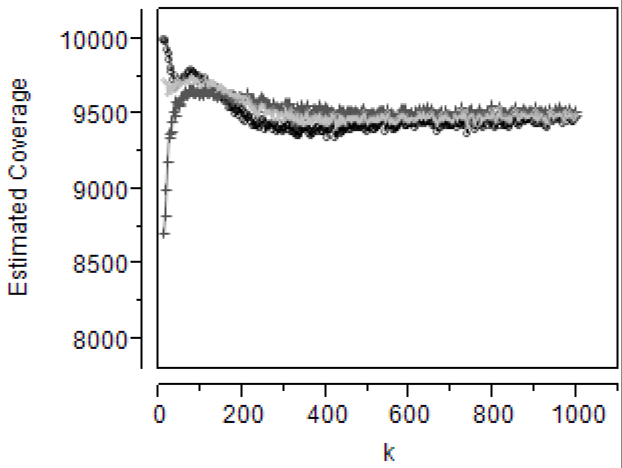
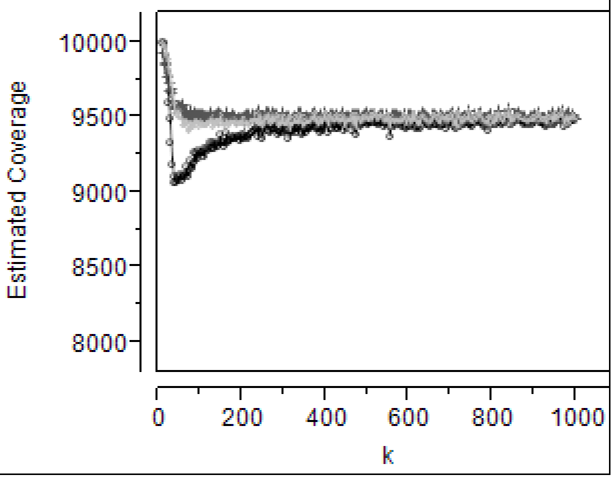


Figure 2: Avg. Width Plots for the Wald, Score, and Likelihood Ratio Confidence Intervals when *n = 0.1k*

From Figure 1, we see that the likelihood ratio interval is conservative (over-covers) or is close to nominal for all the configurations except when $p = 0.50$ and $\phi = 0.05$. However, for this same configuration, the likelihood interval has a coverage that is close to nominal shortly after *k* exceeds 100. The coverage of the score interval appears to converge to the nominal level at a comparable rate to the likelihood ratio interval, whereas the coverage of the Wald interval converges at a much slower rate. In fact, when $p = 0.05$ the Wald interval drastically under-covers, even when *k* is near 1000. The average width results from Figure 2 indicate that the score interval is more narrow for smaller values of *k*. The disparity between the widths of the three intervals seems to go away at a quicker rate for larger values of p and ϕ. It is surprising that the Wald interval is generally the widest, yet it has severe under-coverage problems as indicated in Figure 1. It should also be noted the average widths increase as ϕ increases, which is intuitive because more uncertainty is injected into the estimation problem. Also, not surprisingly, the average widths tend to increase as p approaches 0.50.

Next, we examine the coverage and width properties when $n = 0.4k$ for various combinations of p and ϕ. All simulations were performed in SAS IML V9.3 with 10,000 iterations for each parameter and sample size configuration. Figures 3 and 4 give plots that summarize the coverage and average width properties of confidence intervals (2) - (4). The nominal confidence level is 95% and all estimated coverages have approximate standard errors less than 0.005. The maximum standard error for estimated average widths in figure 4 is 0.0014.

	$\phi = 0.05$	$\phi = 0.25$
$p = 0.05$		
$p = 0.25$		

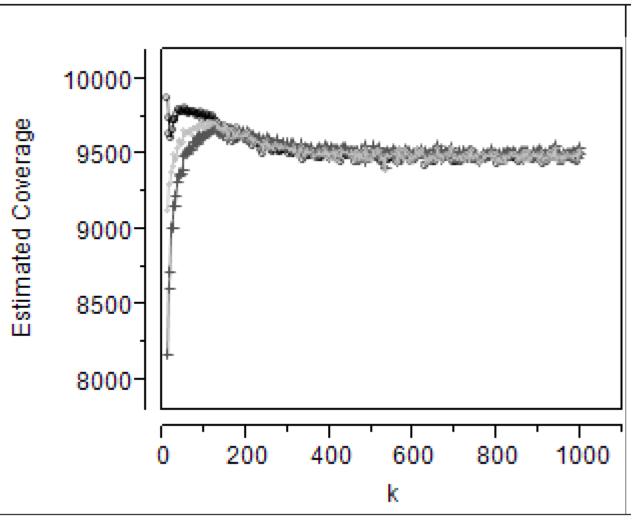
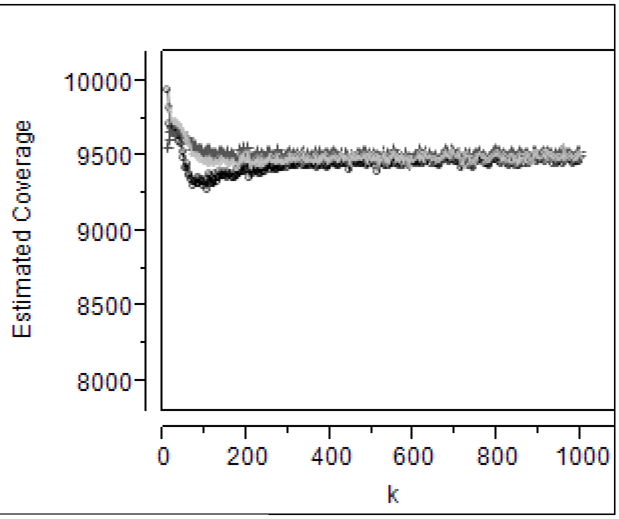
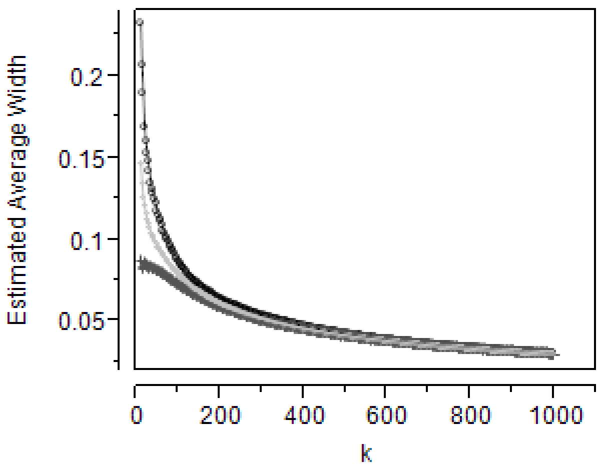
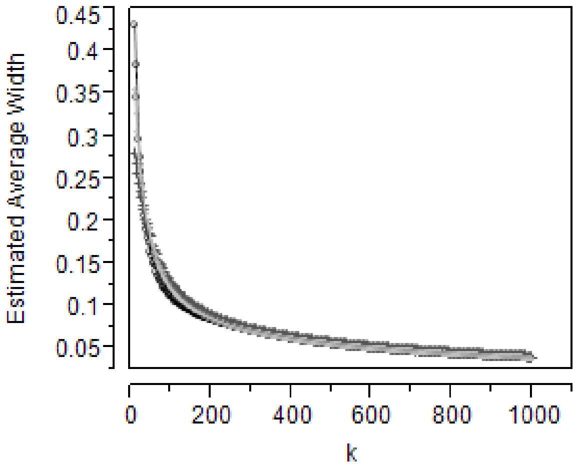
p = 0.50		

Figure 3:Coverage Plots for the Wald, Score, and Likelihood Ratio Confidence Intervals when n = 0.4k

	$\phi = 0.05$	$\phi = 0.25$
p = 0.05		

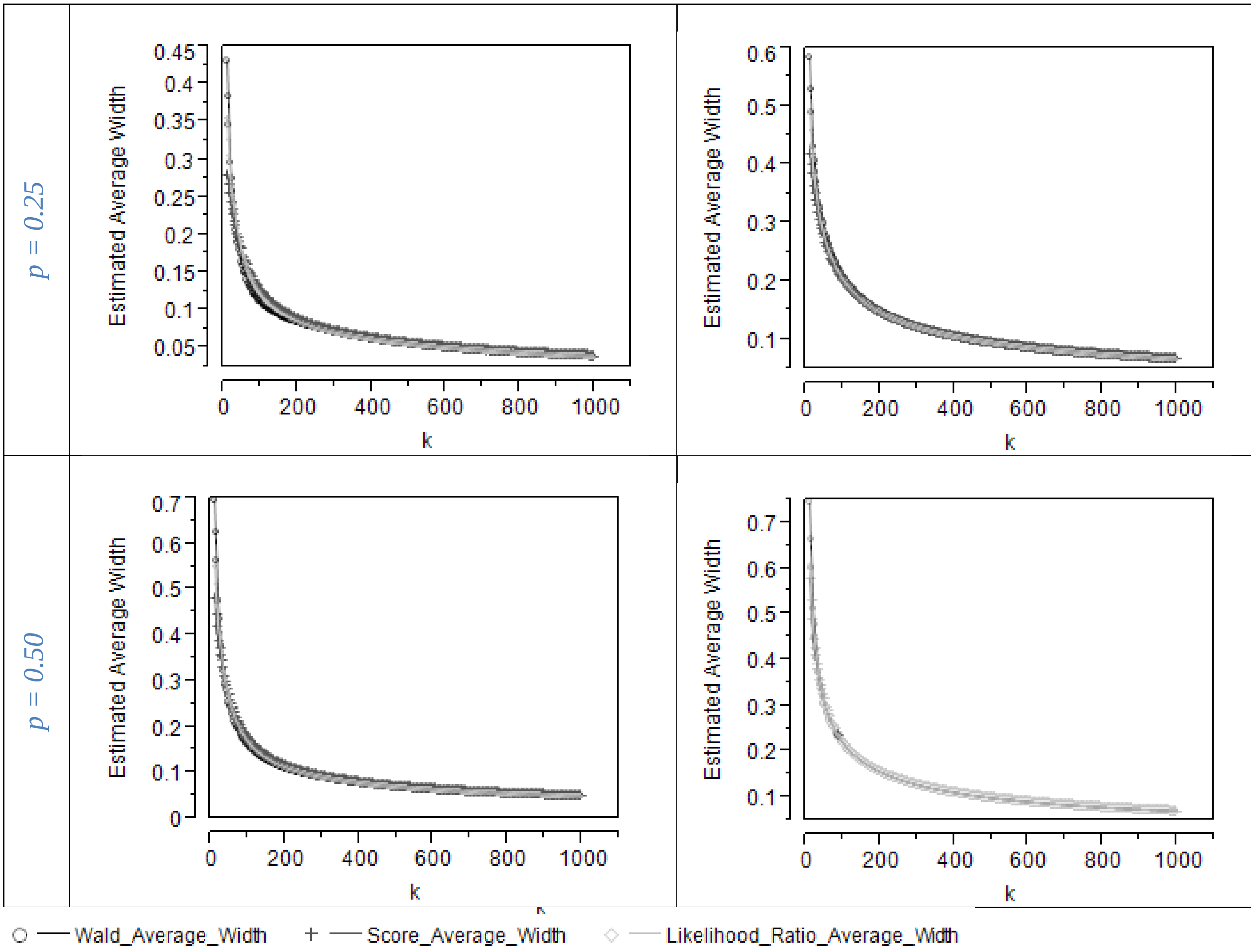


Figure 4:Avg. Width Plots for the Wald, Score, and Likelihood Ratio Confidence Intervals when *n = 0.4k*

From Figure 3, we see that all three intervals generally have better coverage properties as compared to the cases where $n = 0.1k$. These results are not surprising, as we now have more infallible data than in Figure 1. Also, from Figure 4, the three intervals have similar average widths except for the case where $p = 0.05$ and $\phi = 0.05$. In this case, the score interval is narrowest for smaller values of *k*. In all other cases it is not surprising that the three intervals have comparable widths because of the large portion of "good data" we have (i.e. $n = 0.4k$), which translates to the common asymptotic distribution better approximating the Wald, score, and likelihood ratio statistics. The same tendencies that were observed in Figure 2 with an increase in ϕ and p are also present in Figure 4, that is, the average widths tend to increase with an increase in ϕ and p. Also, it should be stated that for the same $k, p,$ and ϕ combination, the widths from Figure 4 are considerably smaller than widths from Figure 2. This gain in precision is intuitive given the increase in the size of the infallible data set.

5. An Application to a Medical Data Set

We now apply the statistical model from Section 2 and three confidence intervals from Section 3 to a medical data set from Hildesheim (1991). Boese (2006) has also considered this data set. The data was generated under fixed sampling, but for illustrative purposes of the model in Section 2, we will assume inverse sampling. This data involves a large study which examines the relationship between the herpes simplex virus and cervical cancer. For us, of interest from the data is estimating the prevalence, p, of the herpes simplex virus in women who have invasive cervical cancer. One diagnostic test for the virus is the western blot procedure (WBP), which is fallible. A more accurate procedure, the refined western blot procedure (RWBP) is accurate and we will consider it as an infallible classifying device. For illustrative purposes we will only consider false positives from the data in stage 2 and allow false negatives to be absorbed into n_{11}. For the two stages, the following counts were observed: $y = 375,\ k = 318,\ n_{00} = 13,\ n_{10} = 3$, and $n_{11} = 23$.

Using the regular inverse sampling model that does not account for misclassification, we get the following point estimate and 95% confidence interval estimate for p : 0.541 and (0.501, .581). Using the more appropriate inverse sampling model that allows for false-positive misclassification, we get the following point estimates: $\hat{p} = 0.485$ and $\hat{\phi} = 0.123$. 95% Wald, score, and likelihood ratio confidence intervals are (0.404, 0.565), (0.397, 0.539), and (0.395, 0.546), respectively. Note that the original inverse sampling model overestimates p, which is not surprising due to the estimated 12.3% false-positive rate. Also, note that the intervals presented here are slightly wider than the ones found in Boese (2006). This is likely due to the fact that Boese (2006) has a slightly smaller estimate (0.119) of the false-positive rate.

Now, one could use only the infallible data to estimate p, but the resulting standard error is higher as compared to the estimator that uses both the fallible and infallible samples. The infallible-only estimate is $\tilde{p} = \frac{n_{11}}{n} = 0.590$, which has a standard error of 0.079. This is considerably higher than the standard error of the estimator (0.041) that uses both the fallible and infallible data.

6. Comments

In this article, we derived an inverse-sampling model that allows for false-positive misclassification and three confidence intervals for the proportion of interest. All three confidence intervals are based on first-order approximations and formed by inverting the Wald, score, and likelihood ratio statistics. Necessary for the development of the confidence intervals was Fisher's information matrix, unrestricted MLE's, and restricted MLE's.

The Wald, score, and likelihood ratio confidence intervals were then studied via a Monte Carlo simulation study. The study indicates that the score and likelihood ratio intervals perform better than the Wald interval in terms of coverage and average width, particularly when the ratio of infallible data to fallible data is small ($n = .1k$). In such cases, we recommend the score or likelihood ratio confidence intervals. However, if the infallible data set is large and the ratio of infallible to fallible data is large, the Wald interval may be preferred as its coverage and width properties are comparable to the score and likelihood ratio intervals, but its computation is easier. Finally, we applied this newly derived model and interval estimators to a real medical data set, which demonstrated the bias present in a model that does not account for misclassification, when, in fact, misclassification is present in the data.

The author acknowledges that this statistical model and resulting estimators are similar to the ones presented in Boese (2006), but with the major distinction of stage 1 in this model using inverse-sampling whereas Boese (2006) uses fixed-sampling. The statistical model and interval estimators presented in this article should prove useful for the practitioner who desires to estimate a particular proportion using inverse-sampling but where the data is subject to false-positive misclassification.

Acknowledgements

This work is supported, in part, by a research grant from the Office of Research and Sponsored Programs at Stephen F. Austin State University.

Appendix

Provided are the second derivatives and negative expectations required for Fisher's information matrix:

$$\frac{\partial^2 \ell}{\partial p^2} = \frac{-n_{00}}{(1-p)^2} - \frac{n_{10}}{(1-p)^2} - \frac{n_{11}}{p^2} - \frac{k(1-\phi)^2}{(p+(1-p)\phi)^2} - \frac{y}{(1-p)^2},$$

$$\frac{\partial^2 \ell}{\partial \phi^2} = \frac{-n_{00}}{(1-\phi)^2} - \frac{n_{10}}{\phi^2} - \frac{k(1-p)^2}{(p+(1-p)\phi)^2} - \frac{y}{(1-\phi)^2},$$

$$\frac{\partial^2 \ell}{\partial p \partial \phi} = \frac{-k}{p+(1-p)\phi} - \frac{k(1-\phi)(1-p)}{(p+(1-p)\phi)^2},$$

$$I_{pp} = \frac{n(1-\phi)}{(1-p)} + \frac{n\phi}{(1-p)} + \frac{n}{p} + \frac{k(1-\phi)^2}{(p+(1-p)\phi)^2} + \frac{k(1-(p+(1-p)\phi))}{(p+(1-p)\phi)(1-p)^2},$$

$$I_{\phi\phi} = \frac{n(1-p)}{(1-\phi)} + \frac{n(1-p)}{\phi} + \frac{k(1-p)^2}{(p+(1-p)\phi)^2} + \frac{k(1-(p+(1-p)\phi))}{(p+(1-p)\phi)(1-\phi)^2},$$

$$I_{p\phi} = I_{\phi p} = \frac{k}{(p+(1-p)\phi)^2}.$$

References

[1] Boese, D., Young, D., Stamey, J. (2006). Confidence intervals for a binomial parameter

[2] based on binary data subject to false-positive misclassification. *Computational Statistics and Data Analysis* **50**: 3369-3385.

[3] Bross, I. (1954). Misclassification in 2 x 2 tables. *Biometrics* **10**: 478-486.

[4] Hildesheim, A., Mann, V., Brinton, L.A., Szklo, M., Reeves, W.C., Rawls, W.E. (1991).

[5] 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.

[6] Lie, R.T., Heuch, I., Irgens, L.M. (1994). Maximum likelihood estimation of the

[7] proportion of congenital malformations using double registration systems. *Biometrics* **50**: 433-444.

[8] Moors, J.J.A., van der Genugten, B.B., Strijbosch, L.W.G. (2000). Repeated audit

[9] controls. *Statistica Neerlandica* **54**: 3-13.

[10] Tenennbien, A. (1970). A double sampling scheme for estimating binomial data with

[11] misclassifications. *Journal of American Statistical Association* **65**: 1350-1361.

[12] Tian, M., Tang, M., Ng, H., and Chan, P. (2009). A comparative study of confidence

[13] intervals for negative binomial proportion. *Journal of Statistical Computation and Simulation* **79**: 241-249.