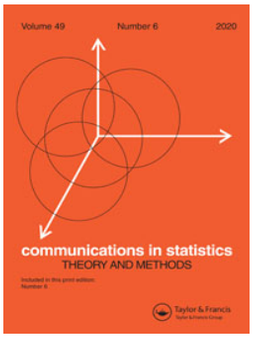

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Prediction intervals for hypergeometric distributions

Kalimuthu Krishnamoorthy & Shanshan Lv

To cite this article: Kalimuthu Krishnamoorthy & Shanshan Lv (2020) Prediction intervals for hypergeometric distributions, Communications in Statistics - Theory and Methods, 49:6, 1528-1536, DOI: 10.1080/03610926.2018.1563181

To link to this article: https://doi.org/10.1080/03610926.2018.1563181

Published online: 23 Jan 2019.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2020, VOL. 49, NO. 6, 1528–1536
https://doi.org/10.1080/03610926.2018.1563181

Prediction intervals for hypergeometric distributions

Kalimuthu Krishnamoorthy and Shanshan Lv

Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA

ABSTRACT

The problem of constructing prediction intervals (PIs) for a future sample from a hypergeometric distribution is addressed. Simple closed-form approximate PIs based on the Wald approach, the joint sampling approach, and a fiducial approach are proposed and compared in terms of coverage probability and precision. Construction of the proposed PIs are illustrated using an example.

ARTICLE HISTORY
Received 23 July 2018
Accepted 10 December 2018

KEYWORDS
Coverage probability; Fiducial approach; precision; Wald method; score method

1. Introduction

Predicting the values of a future random variable on the basis of the past and current samples is an important problem in statistical applications. The prediction problem has been well addressed for various continuous probability models and other parametric models such as linear regression and mixed models. However, compared with the continuous distributions, only limited investigations for discrete distributions are available. Prediction intervals (PIs) for a discrete distribution are used to predict the number of events that may occur in the future. An exact PI for a binomial random variable was proposed by Thatcher (1964). As the exact PI is not simple to compute and too conservative, simple closed-form PI for a binomial random variable is proposed in Nelson (1982) was widely used. Alternative closed-form PIs which are better than the one in Nelson (1982) were proposed in Wang (2010) and Krishnamoorthy and Peng (2011).

In the context of estimating the proportion in a finite population, hypergeometric distribution was used by many authors to develop confidence intervals (CIs). Recently, Wang (2015) have proposed methods of obtaining exact one-sided confidence limits and exact two-sided CIs for the number items with an attribute of interest in a finite lot. He also addressed the problem of interval estimating the finite population size given other parameters and a sample. Young (2015) has obtained one-sided as well as two-sided tolerance intervals for a hypergeometric distribution. However, to the best of our knowledge the prediction problem involving a hypergeometric distribution is never addressed in the literature. One could use the binomial PI for the hypergeometric case provided the population is sufficiently large. Burstein (1975) has noted that the binomial-based results for a hypergeometric distribution could be inaccurate unless the population size is around 5000 or more. The prediction problem that we shall address

CONTACT Kalimuthu Krishnamoorthy krishna@louisiana.edu Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana 70504, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

can be described as follows. Consider a hypergeometric distribution with a lot size N_x and unknown number of defective items M_x. Let X be the number of defective items in a sample of n_x drawn from the lot without replacement. For convenience, we write $X \sim H(n_x, M_x, N_x)$. The probability mass function of X is given by

$$P(X = x|n_x, M_x, N_x) = \frac{\binom{M_x}{x}\binom{N_x - M_x}{n_x - x}}{\binom{N_x}{n_x}}, \quad L_x \le x \le U_x \tag{1}$$

where $L_x = \max\{0, M_x - N_x + n_x\}$ and $U_x = \min\{n_x, M_x\}$. Let $Y \sim H(n_y, M_y, N_y)$ independently of X. For a given number of defective items X in a sample of size n_x, the problem is to find a PI for Y under the assumption that $M_x/N_x = M_y/N_y$. Specifically, for a given confidence level $1-\alpha$, the problem is to find two integer valued functions $L(X, n_x, N_x, n_y, N_y; \alpha)$ and $U(X, n_x, N_x, n_y, N_y; \alpha)$ so that

$$P_{X,Y}(L(X, n_x, N_x, n_y, N_y; \alpha) \le Y \le U(X, n_x, N_x, n_y, N_y; \alpha)) \ge 1-\alpha$$

for all M_x.

In the following section, we describe the PIs based on the Wald method, the joint sampling approach and the generalized fiducial method introduced by Hannig (2009). In Section 3, we compare all three PIs in terms of coverage probability and precision. The methods are illustrated using a real life example in Section 4. Some concluding remarks are given in Section 5.

2. Prediction intervals

To describe various prediction intervals in the sequel, let $p = M_x/N_x = M_y/N_y$ and the finite population correction factor $R_x = \frac{N_x - n_x}{N_x - 1}$ so that $\mathrm{Var}(X) = n_x p(1-p)R_x$ and $\mathrm{Var}(Y) = n_y p(1-p)R_y$, where R_y is R_x with x replaced by y.

2.1. Wald prediction interval

Let $\widehat{p}_x = X/n_x$ and $\widehat{Y}_x = n_y\widehat{p}_x$. The Wald-type CI for p is constructed on the basis of the result that

$$\frac{\widehat{p}_x - p_x}{\sqrt{\widehat{\mathrm{Var}}(\widehat{p}_x)}} = \frac{\widehat{p}_x - p_x}{\sqrt{R_x\widehat{p}_x(1-\widehat{p}_x)/n_x}}$$

has the standard normal distribution for large n_x. Similarly, we can find the Wald-type PI on the basis of the asymptotic result that

$$Z = \frac{\widehat{Y}_x - Y}{\sqrt{\widehat{\mathrm{Var}}(\widehat{Y}_x - Y)}} = \frac{(n_y X - n_x Y)}{\sqrt{\widehat{\mathrm{Var}}(n_y X - n_x Y)}} \sim N(0,1) \tag{2}$$

where $\widehat{\mathrm{Var}}(n_y X - n_x Y) = n_x n_y \widehat{p}_x(1-\widehat{p}_x)(n_y R_x + n_x R_y)$. Let $q_\alpha = z_{1-\alpha/2}$ denote the $100(1-\alpha/2)$ percentile of the standard normal distribution. The Wald PI is formed by

the two roots of the equation $|Z| = q_\alpha$, and they can be expressed as

$$[L_w, U_w] = \widehat{Y}_x \pm \frac{q_\alpha}{\sqrt{n_x}} \sqrt{n_y \widehat{p}_x (1-\widehat{p}_x)(n_y R_x + n_x R_y)} \tag{3}$$

As Y assumes only non negative integers, L_w and U_w can be rounded to their nearest integers to have a proper PI.

2.2. Joint sampling approach

This approach is based on the asymptotic joint sampling distribution of a quantity which is a function of $\widehat{p}_{xy} = (X+Y)/(n_x+n_y)$. This type of approach was used by Brown (1982) to obtain a solution to a statistical calibration problem, and Krishnamoorthy and Peng (2011) used such joint sampling approach to find a PI for the binomial case. To outline this approach, let $\widehat{Y}_{xy} = n_y \widehat{p}_{xy}$. Consider the quantity

$$\frac{\widehat{Y}_{xy} - Y}{\sqrt{\mathrm{Var}\left(\widehat{Y}_{xy} - Y\right)}} = \frac{n_y X - n_x Y}{\sqrt{\mathrm{Var}(n_y X - n_x Y)}} \tag{4}$$

where $\mathrm{Var}(n_y X - n_x Y) = n_x n_y p(1-p)(n_y R_x + n_x R_y)$. By replacing this variance by its estimate $\widetilde{\mathrm{Var}}(n_y X - n_x Y) = n_x n_y \widehat{p}_{xy}(1-\widehat{p}_{xy})(n_y R_x + n_x R_y)$, we can find a PI based on the result that $S = (n_y X - n_x Y)/\sqrt{\widetilde{\mathrm{Var}}(n_y X - n_x Y)} \sim N(0,1)$ asymptotically. Specifically, the PI is determined by the roots (with respect to Y) of the equation $S^2 = q_\alpha^2$, where q_α is as defined in Equation (3). Notice that the equation

$$S^2 = \frac{(n_y X - n_x Y)^2}{\widetilde{\mathrm{Var}}(n_y X - n_x Y)} = q_\alpha^2 \tag{5}$$

is quadratic in Y, and the two roots of this equation form a PI for Y. After some algebraic manipulation and letting $C = (n_y R_X + n_x R_y)/(n_x + n_y)$, the two roots of the Equation (5), denoted by L_j and U_j, can be expressed as

$$L_j = \frac{\widehat{Y}_x\left(1 - \frac{q_\alpha^2 C}{n_x+n_y}\right) + \frac{q_\alpha^2 n_y C}{2n_x} - q_\alpha \left[\left(\frac{R_y}{n_x} + \frac{R_x}{n_y}\right)\widehat{Y}_x\left(n_y - \widehat{Y}_x\right) + \frac{q_\alpha^2 n_x^2 C^2}{4n_x^2}\right]^{1/2}}{1 + \frac{q_\alpha^2 n_y C}{n_x(n_x+n_y)}} \tag{6}$$

and

$$U_j = \frac{\widehat{Y}_x\left(1 - \frac{q_\alpha^2 C}{n_x+n_y}\right) + \frac{q_\alpha^2 n_y C}{2n_x} + q_\alpha \left[\left(\frac{R_y}{n_x} + \frac{R_x}{n_y}\right)\widehat{Y}_x\left(n_y - \widehat{Y}_x\right) + \frac{q_\alpha^2 n_x^2 C^2}{4n_x^2}\right]^{1/2}}{1 + \frac{q_\alpha^2 n_y C}{n_x(n_x+n_y)}} \tag{7}$$

where $\widehat{Y}_x = n_y \widehat{p}_x$ and $q_\alpha^2 = z_{1-\alpha/2}^2$. To make the PI to be integer valued, we shall use $[\lceil L_j \rceil, \lfloor U_j \rfloor]$, where $\lceil x \rceil$ is the ceiling function and $\lfloor x \rfloor$ is the floor function, as a PI for Y, and refer to this PI as JS-PI.

Remark. To tackle the extreme cases of $X=0$ and $X = n_x$, both Wald and the score PIs are computed as follows. When $X=0$ is observed, we compute both PIs using $X=0.5$. We also note that both PIs satisfy a natural property that if $[L, U]$ is a PI for Y based on X, then $[n_y-U, n_y-L]$ is a PI for n_y-Y. To satisfy this natural requirement at the extreme outcomes, the PI when $X = n_x$ is computed as $[n_y-U_{0.5}, n_y-L_{0.5}]$, where $[L_{0.5}, U_{0.5}]$ is the PI at $X=0.5$.

2.3. Fiducial prediction intervals

A fiducial distribution for a parameter is essentially a posterior distribution of the parameter without assuming a prior on the parameter (Efron 1998). There are a few different approaches of finding a fiducial distribution. A fiducial distribution for a parameter can be obtained by inverting a hypothesis test as suggested by Fisher (1935), using a functional model relationship between the statistics and the parameters (Dawid and Stone 1982) or by deducing from a random number generating method (Hannig 2009). A CI for M_x or a PI for a future sample can be obtained from a fiducial distribution of M_x. To obtain a fiducial distribution for M_x, we shall use the general idea of Hannig (2009) and Hannig et al. (2016) to develop a fiducial distribution for M_x, and then use the approach of Wang et al. (2012) to find a fiducial PI. To identify the data generating mechanism in a hypergeometric distribution, we note that x^* is a pseudo random number from the $H(n_x, M_x, N_x)$ distribution if

$$P(X \le x^*-1|n_x, M_x, N_x)<U \le P(X \le x^*|n_x, M_x, N_x)$$

where U is a uniform(0,1) random variable (e.g., see Casella and Berger 2001, 249). Let x be an observed value of $X \sim H(n_x, M_x, N_x)$. For a given x, the fiducial distribution of M_x is implicitly determined by

$$P(X \le x-1|n_x, M_x, N_x)<U \le P(X \le x|n_x, M_x, N_x) \tag{8}$$

where U has a uniform(0, 1) distribution. For a given x, a sample from the fiducial distribution of M_x can be obtained by generating $U_1, \ldots, U_N$ and then finding the values of M_x that satisfy the above inequality for each U_i. A confidence interval for M_x or a prediction interval for a future observation can be obtained based on such fiducial samples.

To obtain a fiducial sample based on (x, n_x, N_x), we first note that the pmf in Equation (1) is valid only when $M_x \in [x, x+N_x-n_x]$. Furthermore, for a given (x, n_x, N_x, U), more than one $M_x \in [x, x+N_x-n_x]$ satisfy the inequality Equation (8). As suggested by Hannig et al. (2016), we can select one of the values of M_x that satisfy the inequality at random. Using these facts, the fiducial distribution of M_x can be obtained empirically by generating uniform(0, 1) random numbers $U_1, \ldots, U_N$ and M_x's as follows. For a generated uniform(0, 1) random number U_i, select one element from the set

$$\left\{M_x : P(X \le x-1|n_x, M_x, N_x)<U_i \le P(X \le x|n_x, M_x, N_x)\right\} \tag{9}$$

at random and refer to the selected element as $M^*_{x_i}$. Then the fiducial sample is given by $\{M^*_{x_1}, \ldots, M^*_{x_N}\}$. Following Wang et al. (2012), a predicting fiducial distribution for a future random variable from $H(n_y, M_y, N_y)$ is determined by the sample $Y^*_1, Y^*_2, \ldots, Y^*_N$, where

$$Y_i^* \sim H\left(n_y, M_{y_i}^*, N_y\right), \quad i = 1, 2, ..., N$$

and $M_{y_i}^* = (M_{x_i}^*/N_x)N_y$. The lower and upper 100α percentiles of the Y_i^*'s is a $100(1-2\alpha)\%$ fiducial prediction interval for $Y \sim H(n_y, M_y, N_y)$. Our numerical investigation indicated that a stable PI for Y can be obtained by using $N = 20,000$ runs or more. That is, for a given (x, n_x, N_x), PIs based on different simulations will be similar if $N = 20,000$ or larger is used. For a given (x, n_x, N_x, n_y, N_y) and a nominal confidence level $1-2\alpha$, the R code given in the Appendix can be used to find a fiducial PI for $Y \sim H(n_y, M_y, N_y)$.

3. Coverage and precision studies

Let $[L(x, n_x, N_x, n_y, N_y; \alpha), U(x, n_x, N_x, n_y, N_y; \alpha)]$ be a $100(1-\alpha)\%$ PI for a future $Y \sim H(n_y, M_y, N_y)$. For some assumed values of M_x, the exact coverage probability of the PI can be evaluated using the following expression.

$$\sum_{x=L_x}^{U_x} \sum_{y=L_y}^{U_y} f(x|n_x, M_x, N_x) f\left(y|n_y, M_y, N_y\right) I_{[L(x,n_x,N_x,n_y,N_y;\alpha), U(x,n_x,N_x,n_y,N_y;\alpha)]}(y) \tag{10}$$

where M_y is the nearest integer to $(M_x/N_x)Ny$ and $I_A(y)$ is the indicator function. The expected width of the PI can be evaluated using the above expression with the indicator function replaced by $U(x, n_x, N_x, n_y, N_y; \alpha) - L(x, n_x, N_x, n_y, N_y; \alpha)$. The coverage probability of a good PI should be close to the nominal level $1-\alpha$.

The coverage probabilities of the Wald PI, PI based on the joint sampling approach (JS-PI) and the fiducial PI are evaluated as a function of $p = M_x/N_x$ and for various combinations of (n_x, N_x, n_y, N_y). The calculated coverage probabilities are plotted in the left panel of Figure 1 and the expected widths corresponding to the same combination of (n_x, N_x, n_y, N_y) are plotted in the right panel of Figure 1. Examination of the plots in Figure 1 clearly indicates that the fiducial PI is in general conservative having coverage probabilities greater than or equal to the nominal level.95 in all cases. The Wald PI could be very liberal if n_x is smaller than n_y (see the plots for $(n_x, N_x, n_y, N_y) = (6, 300, 10, 300)$ and $(10, 200, 50, 200)$) and conservative if n_x is larger than n_y (see the plots for $(n_x, N_x, n_y, N_y) = (30, 200, 6, 400)$ and $(50, 200, 10, 200)$). We also notice that all three PIs are too conservative when $p = M_x/N_x$ is near boundaries. On the basis of all six coverage plots, we see that the JS-PI is the only PI which is neither too conservative nor too liberal, having coverage probabilities closer to the nominal level than other two PIs.

Regarding precisions of the PIs, we see from the plots on the right column of Figure 1 that the expected widths of the JS-PI and the fiducial PI are in agreement with their coverage properties. The expected widths of these two PIs are approximately the same when their coverage probabilities are approximately the same; see the plots for $(n_x, N_x, n_y, N_y) = (30, 500, 50, 500)$. In other situations where the fiducial CI is conservative, it is wider than the JS-PI when $p \in [.25, .75]$ for some cases and for all p in other cases; for example, see the plot for $(n_x, N_x, n_y, N_y) = (6, 300, 10, 300)$, (10,300,10,300) and (30,200,6,400). The Wald PI appears to be shorter than the other two PIs when p is near boundary. This Wald PI has a peculiar property; for some values of p, it is too liberal

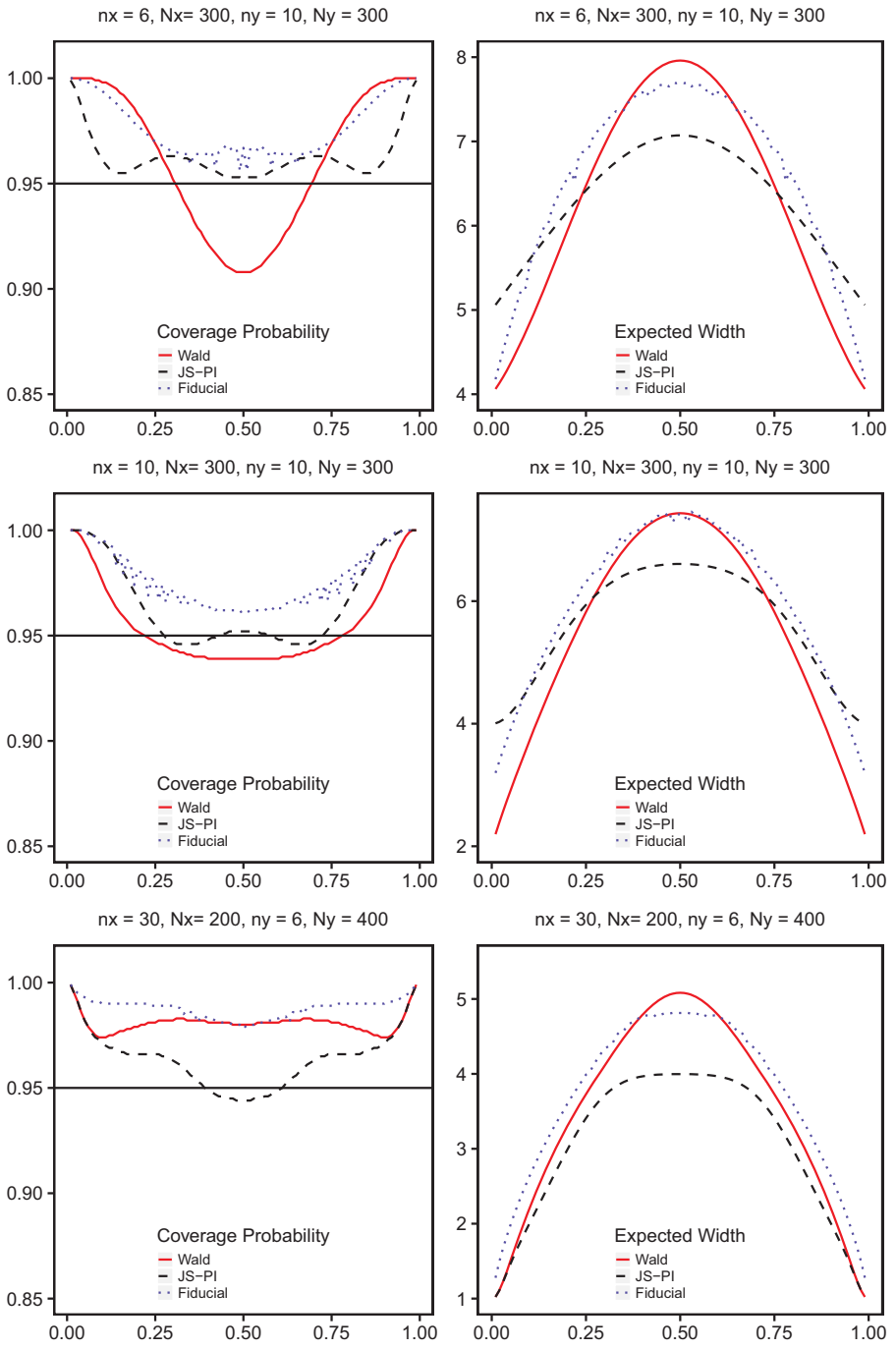

Figure 1. Coverage probabilities and expected widths 95% prediction intervals as a function of $p = M_x/N_x$.

having coverage probabilities appreciably smaller than the nominal level, but its expected widths for the same values of p are larger than those of other two PIs; see the plots for $(n_x, N_x, n_y, N_y) = (6, 300, 10, 300)$ and (10, 200, 50, 200).

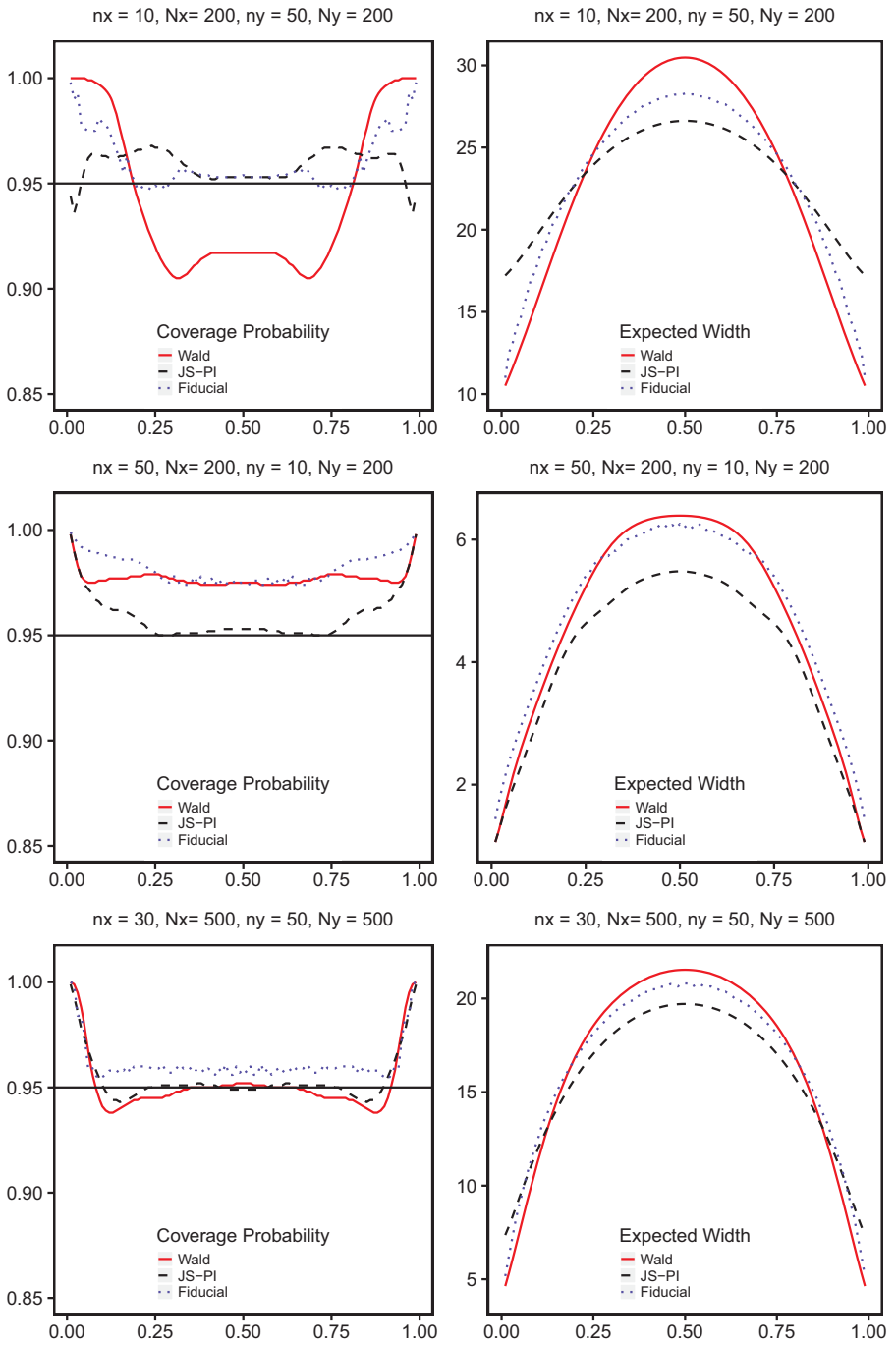

Figure 1. Continued.

On an overall basis, the JS-PI is the one having coverage probabilities close to the nominal level with shorter expected widths in most cases. The fiducial PI appears to guarantee coverage probabilities in most cases.

4. An example

To illustrate the prediction interval methods in the preceding sections, we shall adapt the example in Krishnamoorthy and Thomson (2002). This example involves the problem of estimating/predicting the the number of unacceptable cans produced by canning machines which are identical with respect to manufacturer and model. A can is determined to be unacceptable (for sale) if the content of the can weighs less than 95% of the labeled weight. Inspection of a sample $n_x = 20$ cans from a lot of $N_x = 200$ cans produced by Machine 1 revealed $x = 2$ unacceptable cans.

Suppose it is desired to predict the number of unacceptable cans Y in a future sample of size $n_y = 40$ cans from a lot of $N_y = 500$ cans produced by Machine 2. For this case, the 95% Wald PI is [0, 10] and the 95% JS-PI is [1, 12] and the 95% fiducial PI is [0, 13]. Notice that the Wald PI could be very liberal and so it produced a shorter PI whereas the fiducial method is conservative and produced a PI that is wider than the other two PIs.

5. Summary and conclusions

In this article, we have investigated the PIs for a future random variable from a hypergeometric distribution based on the Wald approach, the joint sampling approach and the fiducial approach. For interval estimation of parameters of discrete distributions, it is known that the Wald approach produces confidence intervals that are too liberal even for somewhat large samples (Brown, Cai, and DasGupta, 2001). On the other hand, the joint sampling approach produces PIs that are very satisfactory in terms of coverage probability and precision. Even though calculation of the fiducial PIs is somewhat involved, it appears to guarantee coverage probabilities for all cases. Our R code can be used to calculate fiducial PI in an easy manner. We also note that this fiducial PI is too conservative at the boundaries like other interval estimates for discrete distributions such as the exact CI for a binomial proportion (see Agresti and Coull 1988). Nevertheless, an exact PI with the coverage probabilities very close to the nominal level is still desirable.

In general, fiducial PI maybe recommended for applications if the coverage requirement is important. The JS-PI is recommended for simplicity and precision despite being a little liberal in some cases.

Acknowledgements

The authors are grateful to a reviewer for providing useful comments and suggestions.

References

Agresti, A., and B. A. Coull. 1988. Approximate is better than exact for interval estimation of binomial proportion. *The American Statistician* 52:119–25.

Brown, P. J. 1982. Multivariate calibration. *Journal of the Royal Statistical Society, Series B* 44: 287–321.

Brown, L. D., T. Cai, and A. DasGupta. 2001. Interval estimation for a binomial proportion (with discussion). *Statistical Science* 16:101–33.

Burstein, H. 1975. Finite population correction for binomial confidence limits. *Journal of the American Statistical Association* 70 (349):67–9.
Casella, G., and R. L. Berger. 2001. *Statistical Inference*. Boston, MA: Cengage Learning.
Dawid, A. P., and M. Stone. 1982. The functional-model basis of fiducial inference. *Annals of Statistics* 10 (4):1054–74.
Efron, B. 1998. R. A. Fisher in the 21st century. *Statistical Science* 13:95–122.
Fisher, R. A. 1935. The fiducial argument in statistical inference. *Annals of Eugenics* VI:91–8.
Hannig, J. 2009. On generalized fiducial inference. *Statistica Sinica* 19:491–544.
Hannig, J., H. Iyer, R. C. S. Lai, and T. Lee. 2016. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association* 111 (515):1346–61.
Krishnamoorthy, K., and J. Peng. 2011. Improved closed-form prediction intervals for binomial and Poisson distributions. *Journal of Statistical Planning and Inference* 141 (5):1709–18.
Krishnamoorthy, K., and J. Thomson. 2002. Hypothesis testing about proportions in two finite populations. *The American Statistician* 56 (3):215–22.
Nelson, W. 1982. *Applied life data analysis*. New York: Wiley.
Thatcher, A. R. 1964. Relationships between Bayesian and confidence limits for prediction. *Journal of the Royal Statistical Society, Series B* 26:176–92.
Wang, C. M., J. Hannig, and H. K. Iyer. 2012. Fiducial prediction intervals. *Journal of Statistical Planning and Inference* 142 (7):1980–90.
Wang, H. 2010. Closed form prediction intervals applied for disease counts. *The American Statistician* 64 (3):250–6.
Wang, W. 2015. Exact optimal confidence intervals for hypergeometric parameters. *Journal of the American Statistical Association* 110 (512):1491–9.
Young, D. S. 2015. Tolerance intervals for hypergeometric and negative hypergeometric variables. *Sankhya Series B* 77 (1):114–40.

Appendix

R code for computing the fiducial prediction intervals based on (x, n_x, N_x, n_y, N_y) and the nominal confidence level $1-2\alpha$.

```
R code
-----------------------------------------------------------
mn=x; mx=x+Nx-nx; Mxu=c()
sq=seq(mn, mx,1) # support of the fiducial distribution of Mx
ps0=phyper(x-1, sq, Nx-sq, nx); ps1=phyper(x, sq, Nx-sq, nx)
u=runif(N) # N=number of uniform variates
for(j in 1:N){
ind=which(ps0<u[j] & u[j] <= ps1)
if(length(ind) == 1){
Mxu[j] = sq[ind]}else{
Mxu[j] = sample(sq[ind],1)}}
Myu = (Mxu/Nx)*Ny
PI=quantile(rhyper(N, Myu, Ny-Myu, ny), c(alpha,1-alpha))
```