01 Jan 1993

Prediction Intervals Based On Partial Observations For Some Discrete Distributions

Lee J. Bain
Missouri University of Science and Technology, ljbain@mst.edu

Jagdish K. Patel
Missouri University of Science and Technology

Recommended Citation
L. J. Bain and J. K. Patel, "Prediction Intervals Based On Partial Observations For Some Discrete Distributions," *IEEE Transactions on Reliability*, vol. 42, no. 3, pp. 459 - 463, Institute of Electrical and Electronics Engineers, Jan 1993.
The definitive version is available at https://doi.org/10.1109/24.257831

Prediction Intervals Based on Partial Observations for Some Discrete Distributions

Lee J. Bain
University of Missouri, Rolla
Jagdish K. Patel
University of Missouri, Rolla

Key Words — **Prediction, Binomial, Hypergeometric, Negative-hypergeometric, Poisson, Negative-binomial, Sufficient statistic**

Reader Aids —
General Purpose: Widen state of art
Special math needed for explanations: Statistics
Special math needed for results: Same
Results useful to: Practicing statisticians

Summary & Conclusions — **Prediction limits of the following type are considered for the binomial, hypergeometric, and negative-binomial distributions. For the binomial distribution, suppose X_r successes have occured in the first r trials, and based on this partial information it is desired to predict the total number of successes X_s ($r < s$), which will have occurred by trial s. Similar results are considered for the Poisson and negative-binomial distributions. The results are expressed in terms of well known distributions which have been tabulated, but it is quite tedious to carry out the procedures based on available tables. Normal (Gaussian) approximations are provided which makes these methods convenient to apply.**

1. INTRODUCTION

Patel & Samaranayake [6] obtained small sample simultaneous conservative prediction intervals for binomial, Poisson, hypergeometric, and negative-binomial distributions. These intervals used an observation X from a past sample to predict observations $Y_i (i = 1,2,...k)$ from k future s-independent samples from the same population. These were obtained using a procedure that connected the prediction interval problem to that which arises in the ranking and selection of statistical populations. This made it possible to use known tables from the literature to obtain prediction factors for these intervals.

For continuous distributions, a standard method to construct prediction intervals is to use appropriate pivotal-quantities [5]. This is not possible for discrete distributions since such pivotal-quantities are not available. As a result not much work on prediction has been done for discrete distributions.

This paper considers prediction intervals of the following type for binomial, Poisson, and negative-binomial distributions. Consider a process which generates a sequence of outcomes, and let X_r be the observed total of all outcomes at stage r. Using this partial information, we would like to predict the future cumulative total X_s of all outcomes at stage s ($r < s$). In the binomial case, if we are planning to conduct s trials, then X_r & X_s are the total number of successes observed by trials r & s, respectively. For the Poisson case, if we are planning to observe the process during the time interval $[0,s]$, then X_r & X_s are the number of events observed during the time intervals $[0,r]$ & $[0,s]$, respectively. For the negative-binomial case, if we are planning a sequence of trials to observe success s, then X_r & X_s are the total number of trials needed to observe successes r & s, respectively. In these cases r.v.'s X_r & $(X_s - X_r)$ statistically independent, although r.v.'s X_r & X_s are not.

Prediction intervals for discrete distributions are useful in many applications. For example, in life testing and reliability studies, consider a complex piece of equipment which is observed continuously, and the total number of breakdowns per week is recorded. Suppose by week $r = 40$, there are $X_r = 6$ breakdowns of the equipment. Using this partial information, we want to know the total number X_{100} of all breakdowns which will have occured by week $s = 100$.

Or, in the negative binomial setting, we wish to have a prediction interval on the time, in weeks, that will be achieved before breakdown 12 occurs. That is how many trouble-free weeks will occur before non-trouble-free week 12 occurs. Large-sample approximations are derived in sections 3-5.

Notation

binf$(x;p,N)$ $\quad \sum_{i=0}^{x} \frac{N!}{i!(N-i)!} p^i (1-p)^{N-i}$, for $x=0,1,\ldots,N$

poif$(x;\mu)$ $\quad \sum_{i=0}^{x} \exp(-\mu)\mu^i/i!$, for $x=0,1,2,\ldots$

hypf$(x;s,t,r)$ $\quad \sum_{i=0}^{x} \frac{r!}{i!(r-i)!} \frac{(s-r)!}{(t-i)!(s-r-t+i)!} \Big/ \frac{s!}{t!(s-t)!}$, for $\max(0,t-s+r) \le x \le \min(t,r)$

nbif$(x;p,k)$ $\quad \sum_{i=k}^{x} \frac{(i-1)!}{(k-1)!(i-k)!} p^k (1-p)^{i-k}$, for $x=k,k+1,\ldots$

nhyf$(x;t-1,s-1,r)$ $\quad \sum_{i=r}^{x} \frac{(i-1)!}{(r-1)!(i-r)!} \frac{(t-i-1)!}{(s-r-1)!(t-i-s+r)!} \Big/ \frac{(t-1)!}{(s-1)!(t-s)!}$, for $x=r, r+1,\ldots,t-s+r$

$X_r = X$ $\quad$ observed total of all outcomes at stage r

0018-9529/93/$3.00 ©1993 IEEE

$X_s = T$	future total of all outcomes at stage s $(r<s)$
$F_{X\|t}(x)$	$\Pr\{X \le x \| T = t\}$
$1-\alpha$	Pr{a random interval based on X_r contains the future observation X_s}
$F(x;\theta)$	Cdf of X_r
Z_δ	lower δ quantile of the standard s-normal distribution.

Other, standard notation is given in "Information for Readers & Authors" at the rear of each issue.

2. METHOD

Upper and lower prediction limits for X_s are obtained by considering the conditional distribution of X_r given X_s, to eliminate the unknown parameter, and then solving implicitly for the limits, as is done in the general method of constructing s confidence intervals when pivotal quantities are not available [1]. This approach is analogous to one discussed by Faulkenberry [2] which he applied to the Poisson distribution, where he considers the conditional distribution of the future variable, given a s-sufficient statistic. The present form provided a more direct solution, and a more direct derivation of large-sample results. For convenience, let $X_r = X$, $X_s - X_r = Y$, $X_s = T = X + Y$. Let T be a complete s sufficient statistic for θ. The following conditional probability

$$\Pr\{h_1(t) < X < h_2(t) | T = t\}, \tag{1}$$

with functions h_1 and h_2 to be selected later, does not depend on θ. We rewrite this conditional probability as

$$F_{X|t}[h_2(t) - 1] - F_{X|t}[h_1(t)]. \tag{2}$$

For a given t, suppose $h_2(t)$ is an increasing function of t, so that

$$1 - F_{X|t}[h_2(t) - 1] = \alpha_1, \tag{3}$$

and let $t_L(\alpha_1)$ be the solution of $h_2(t_L(\alpha_1)) = x_0$, where x_0 is the observed value of X. Then the computed value, $t_L(\alpha_1) \le t$ iff $x_0 \le h_2(t)$. Thus a $(1-\alpha_1)$ lower prediction interval $[t_L(\alpha_1), \infty]$ for X_s is obtained by setting $h_2(t) = x_0$ in (3) and solving for $t = t_L(\alpha_1)$.

Similarly, for a given t, suppose $h_1(t)$ is an increasing function of t so that

$$F_{X|t}[h_1(t)] = \alpha_2 \tag{4}$$

and let $t_U(x)$ be the solution of $h_1(t_U(\alpha_2)) = x_0$. Then the computed $t_U(\alpha_2) \ge t$ iff $x_0 \ge h_1(t)$. Thus a $(1-\alpha_2)$ upper prediction interval $[0, t_U(\alpha_2)]$ for X_s is obtained by setting $h_1(t) = x_0$ in (4) and then solving it for $t = t_U(\alpha_2)$.

If $\alpha_1 + \alpha_2 = \alpha$, a $(1-\alpha)$ 2-sided prediction interval for X_s is obtained as $[t_L(\alpha_1), t_U(\alpha_2)]$ by setting $h_1(t) = x_0$ and $h_2(t) = x_0$ in (3) & (4), respectively and then solving them for t.

Although these prediction limits are determined from the conditional distribution of X given $T = t$, as usual, they are also unconditionally $(1-\alpha)$ level limits as well;

$$\Pr\{h_i(T) < X < h_2(T)\}$$
$$= \mathrm{E_T}\{\Pr\{h_1(t) < X < h_2(t) | T = t\}\}$$
$$= \mathrm{E_T}\{1-\alpha\} = 1-\alpha. \tag{5}$$

3. BINOMIAL DISTRIBUTION

For $1 \le r < s$, let X_r & X_s be two binomial r.v.'s with Cdf's $\mathrm{binf}(x;p,r)$ and $\mathrm{binf}(x+y;p,s)$, respectively. Then it follows from the s-independence of the Bernoulli trials associated with this distribution that the r.v. $(X_s - X_r)$ is s-independent of the r.v. X_r and has Cdf $\mathrm{binf}(y;p,s-r)$. The statistic X_s is known to be s-complete and s-sufficient for p. We first consider the following known result.

$$\Pr\{X=x|T=t\} = \Pr\{X=x,\ T=t\}/\Pr\{T=t\}$$
$$= \Pr\{X=x,\ Y=t-x\}/\Pr\{T=t\}$$
$$= \frac{\binom{r}{x}p^x(1-p)^{r-x}\binom{s-r}{t-x}p^{t-x}(1-p)^{s-r-t+x}}{\binom{s}{t}p^t(1-p)^{s-t}}$$
$$= \binom{r}{x}\binom{s-r}{t-x}/\binom{s}{t}, \tag{6}$$

which is a hypergeometric distribution. Then by the method in section 2, $t_L(\alpha_1)$ and $t_U(\alpha_2)$ are solutions of t obtained from:

$$\mathrm{hypf}(x_0-1;s,t,r) = 1-\alpha_1 \tag{7}$$

$$\mathrm{hypf}(x_0;s,t,r) = \alpha_2, \tag{8}$$

respectively. Extensive tables of the Cdf are available in [4]. In their notation, $N=s$, $n=t$, $k=r$, and $x=x_0$. In some cases it is necessary to use symmetry properties of the hypergeometric distribution [4:pp 4-5]. These tables cover the distribution for $N = 1(1)49, 50(10)100$ and 1000. The values for $N = 1000$ are given only for $n = 500$. Some values for $N = 100(100)2000$ are also given.

Use of these tables can be quite tedious. Of course exact solutions can be obtained with the aid of a computer as well, but the following s-normal approximation provides a simple convenient solution which is adequate even for small samples. The larger sample size is the most important anyway, since by the nature of prediction limits it is usually not possible to obtain tight bounds based on few data, even though an efficient method is used. However, knowledge that a wide range of possible outcomes can reasonably occur, is, in itself, useful.

The approximate prediction limits $t_L(\alpha_1)$ and $t_U(\alpha_2)$ using the s-normal approximation are obtained as follows. For large r-value, consider:

$\Pr\{X_r \leq h_2(t) - 1\}$

$$\approx \Pr\left\{Z \leq \frac{h_2(t)-1-E(X_r) + 1/2}{\sqrt{\mathrm{Var}(X_r)}}\right\} = 1-\alpha_1 \quad (9)$$

$$E\{X_r\} = t(r/s),$$

$$\mathrm{Var}\{X_r\} = tr(s-r)(s-t)/[s^2(s-1)].$$

To solve for $t_L(\alpha_1)$ we consider:

$$h_2(t) - 1/2 - t(r/s) = z_{1-\alpha_1}\sqrt{\frac{tr(s-r)(s-t)}{s^2(s-1)}}. \quad (10)$$

Set $X_0 = h_2(t)$ and square both sides of (10).

$$t_L(\alpha_1) \approx \frac{(2x_0'v+sw_1) - \sqrt{s^2w_1^2 + 4x_0'w_1(r-x_0')}}{2(v^2 + w_1).} \quad (11)$$

$v \equiv (r/s)$, $w_1 \equiv z_{1-\alpha_1}^2 v(1-v)/(s-1)$, $x_0' \equiv x_0 - ½$

In an analagous way —

$$t_U(\alpha_2) \approx \frac{(2x_0''v+sw_2) + \sqrt{s^2w_2^2 + 4x_0''w_2(r-x_0'')}}{2(v^2 + w_2)} \quad (12)$$

$w_2 = z_{1-\alpha_2}^2 v(1-v)/(s-1)$, $x_0'' = x_0 + ½$

For example, suppose for a total random sample of size $s=500$, $r=100$ devices are tested and $x_0=5$ are unacceptable. A 90% prediction interval is desired for the total number of unacceptable devices, x_s, in the lot of 500. In this case $\alpha_1=\alpha_2=0.05$, $v= 0.2$, $x_0'=4.5$, $x_0''=5.5$, $w_1s=w_2s=500$ $(1.645)^2$ $(0.2)(0.8)/499=0.4338$, and (11) and (12) give:

$$[t_L(0.05), t_U(0.05)] = [11.47, 49.48] \approx [11,49].$$

The prediction interval for Y, the number of unacceptable items in the remaining $s-r=400$ items, is:

$$[y_L,y_U] = [t_L-x_0, t_U-x_0] = [6,45] \quad (13)$$

The true probability that an item is unacceptable, p, is unknown, and that is taken into account in the above prediction intervals. To illustrate the effect due to p being unknown, if p is known with $p=5/100=0.05$, then a 90% prediction interval for x_s is [17,33], since

$$\Pr\{17 \leq X_s \leq 33\} = \mathrm{binf}(33;\ 0.05,\ 500)$$

$$-\ \mathrm{binf}(16;\ .05,\ 500) = 0.955-0.034 = 0.921.$$

(The conservative level, 0.921, is obtained due to the discreteness of the variables).

Similarly, if $\hat{p}=0.05$, is obtained based on the first 100, and then a prediction interval is computed for the number of unacceptable in the next 400, using $p=0.05$ as if it were a known value, one obtains the prediction interval [13,27], since

$$\mathrm{binf}(27;\ 0.05,\ 400) - \mathrm{binf}(12;\ 0.05,\ 400)$$

$$= 0.952-0.036 = 0.916.$$

For p unknown, the proper prediction interval in this case is [6,45], as shown in (13).

Of course, the loss due to not knowing p is reduced if more data are available. Suppose $x_0=20$ is observed from a sample of $r=400$, and a prediction interval is desired for the total, X_s, of unacceptable items in $s=500$ items. In this case (11) and (12) give the 90% prediction interval $[t_L, t_U] = [20.7,30.0] \approx [21,30]$, and $[y_L,y_U] = [1,10]$.

If $p=0.05$ had been considered known in this case the corresponding prediction interval for the number of unacceptable devices in the last 100 items becomes [2,9], where $\mathrm{binf}(9;0.05,100)-\mathrm{binf}(1,0.05,100) = 0.972-0.037 = 0.935$.

The prediction limits above were computed using the s-normal approximation. If the hypergeometric cumulative is used directly, (7) & (8), the same prediction interval, [11,49], is obtained, since

$$\mathrm{hypgf}(5;500,49,100) = 0.046,\ \mathrm{hypgf}(4;500,11,100) = 0.952$$

For the prediction interval [21,30], the exact probability level is

$$\mathrm{hypf}(19;500,21,400) - \mathrm{hypf}(20;500,30,400)$$

$$= 0.946 - 0.055 = 0.891$$

4. POISSON DISTRIBUTION

For $0<r<s$, let X_r & X_s be two Poisson r.v.'s which count the number of events occuring in time periods $[0,r]$ and $[0,s]$, respectively. It follows from the s-independence property of the process associated with this distribution that the r.v. (X_s-X_r) is s-independent of the r.v. X_r and its Cdf is $\mathrm{poif}(y;(s-r)\lambda)$. The statistic X_s is known to be s-complete and s-sufficient for λ. We first consider the following result.

$$\Pr\{X=x|T=t\} = \Pr\{X=x,\ T=t\}/\Pr\{T=t\}$$

$$= \Pr\{X=x,\ Y=t-x\}/\Pr\{T=t\}$$

$$= \binom{t}{x}\left(\frac{r}{s}\right)^x\left(1-\frac{r}{s}\right)^{t-x} = \mathrm{binm}(x;r/s,t).$$

By the method discussed in (3) & (4). $t_L(\alpha_1)$ and $t_U(\alpha_2)$ would be solutions of t obtained from:

$$\mathrm{binf}(x_0-1;r/s,t) = 1-\alpha_1 \tag{14}$$

$$\mathrm{binf}(x_0;r/s,t) = \alpha_2, \tag{15}$$

respectively. Binomial distribution tables are widely available, and convenient approximate prediction limits, $t_L(\alpha_1)$ & $t_U(\alpha_2)$, using a s-normal approximation are obtained, as in the binomial case, with $\mathrm{E}\{X_r\} = t(r/s)$, and $\mathrm{Var}\{X_r\} = t(r/s)(1-r/s)$.

$$t_L(\alpha_1) = \frac{(2x_0'+w_1) - \sqrt{w_1(w_1+4x_0')}}{2v} \tag{16}$$

$$t_U(\alpha_2) = \frac{(2x_0''+w_2) + \sqrt{w_2(w_2+4x_0'')}}{2v} \tag{17}$$

$$w_1 = z_{1-\alpha_1}^2(1-v),\ w_2 = z_{1-\alpha_2}^2(1-v).$$

Suppose in a fatigue testing experiment, a component is replaced when it fails, and the experiment continues. Suppose $x_0=5$ failures were observed during $r=100$ stress cycles (or weeks, months, etc), and a 90% prediction interval is desired for the total number of failures, X_s, which will have occurred after $s=500$ stress cycles. We have $x_0'=4.5$, $x_0''=5.5$, $v=0.2$, $w_1=w_2=2.165$, and $[t_L,t_U]=[11.4,51.0]\approx[11,51]$.

The exact probability level for the prediction interval $[11,51]$, from (15) and (16) is $\mathrm{binf}(4;0.2,11)-\mathrm{binf}(5;0.2,51) = 0.950-0.042 = 0.908$.

5. NEGATIVE-BINOMIAL

For $1\le r<s$, let X_r &X_s be two negative-binomial r.v.'s which count the number of trials needed to obtain successes r & s, respectively. It follows from the s-independence of the trials, that the r.v. (X_s-X_r) is s-independent of the r.v. X_r and has Cdf, $\mathrm{nbif}(y;p,s-r)$. The statistic X_s is known to be s-complete and s-sufficient for p. We first consider the following result.

$$\Pr\{X=x|T=t\} = \Pr\{X=x,\ T=t\}/\ \Pr\{T=t\}$$

$$= \Pr\{X=x,\ Y=t-x\}/\Pr\{T=t\}$$

$$= \frac{\binom{x-1}{r-1}p^r(1-p)^{x-r}\cdot\binom{t-x-1}{s-r-1}p^{s-r}(1-p)^{t-x-s+r}}{\binom{t-1}{s-1}p^s(1-p)^{t-s}}$$

$$= \binom{x-1}{r-1}\binom{t-x-1}{s-r-1}\Big/\binom{t-1}{s-1}, \tag{18}$$

which is an inverse (negative) hypergeometric distribution [3]. Let $\mathrm{nhyf}(x;\ t-1,s-1,r)$ denote its Cdf. Then, $t_L(\alpha_1)$ and $t_U(\alpha_2)$ are solutions of t obtained from:

$$\mathrm{nhyf}(x_0-1;\ t-1,s-1,r) = 1-\alpha_1 \tag{19}$$

$$\mathrm{nhyf}(x_0;\ t-1,s-1,r) = \alpha_2, \tag{20}$$

respectively. Inverse hypergeometric probabilities can be obtained from hypergeometric probabilities by using:

$$\mathrm{nhyf}(x;\ t-1,s-1,r) = 1-\mathrm{hypf}(r-1;\ t-1,x,s-1). \tag{21}$$

Because of (21), no new tables are needed.

The approximate prediction limits $t_L(\alpha_1)$ and $t_U(\alpha_2)$ using the s-normal approximation for large r are obtained as in previous cases with $\mathrm{E}\{X_r\} = t(r/s)$, and $\mathrm{Var}\{X_r\} = rt(t-s)(s-r)/[s^2(s+1)]$.

$$t_L(\alpha_1) = \frac{(2x_0'v-sw_1) - \sqrt{s^2w_1^2 - 4x_0'w_1(r-x_0')}}{2(v^2 - w_1)} \tag{22}$$

$$t_U(\alpha_2) = \frac{(2x_0''v-sw_2) + \sqrt{s^2w_2^2 - 4x_0''w_2(r-x_0'')}}{2(v^2 - w_2).} \tag{23}$$

$$w_1 \equiv z_{1-\alpha_1}^2 v(1-v)/(s+1),\ w_2 \equiv z_{1-\alpha_2}^2 v(1-v)/(s+1).$$

We illustrate these prediction intervals with the following example (with some changes) from [7]. Suppose a machine produces parts which are 100% inspected. It is of interest to find the number of parts which need to be inspected to find nonconforming unit 8 in the inspection program. This is when the process is halted and a maintenance person is called in to reset the machine. Now suppose 11 parts were inspected by the time nonconforming part 4 was obtained and that we are interested in predicting X_8. Here $r=4$, $s=8$, $x_0=11$. Let $\alpha_1=\alpha_2=0.05$. We solve $\mathrm{nhyf}(10;\ t-1,7,4) = 1-\mathrm{hypf}(3;\ t-1,10,7) = 1-0.035 \ge 0.95$ for $t=t_L(0.05)=15$ using (19) and (21) and solve $\mathrm{nhyf}(11;\ t-1,7,4) = 1-\mathrm{hypf}(3;\ t-1,11,7) = 1-0.9504 \le 0.05$ for $t = t_U(0.50)=46$ using (20) and (21). Thus a 90% 2-sided equal-tail prediction interval for X_s, based on $X_r=11$, is $[15,46]$. The actual prediction probability is $(0.9650-0.0496) = .9154$.

For the approximate limits, we use (22) and (23) and calculate $v = (4/8)=0.50$, $x_0' = 10.5$, $x_0'' = 11.5$, $w_1=w_2 =(1.645)^2\ (0.5)(0.5)/9 = 0.0752$ to obtain

$$t_L(0.50) \approx \frac{9.899 - 4.570}{0.350} = 15.22$$

$$t_L(0.50) \approx \frac{10.898 + 5.128}{0.350} = 45.79$$

So an approximate 90% 2-sided equal-tail prediction interval for X_s, based on $X_r = 11$ is $[15,46]$.

REFERENCES

[1] L. J. Bain, M. Engelhardt, *Introduction to Probability and Mathematical Statistics*, 1989; PWS-Kent.

[2] G. D. Faulkenberry, "A method of obtaining prediction intervals". *J. Amer Statistical Assoc.*, vol 68, 1973, pp 433–435.

[3] W. C. Guenther, Hypergeometric Distributions, *Encyclopedia of Statistical Sciences*, vol 3 (editors. S. Kotz, N. L. Johnson, C. B. Read), 1981, pp 707–712; John Wiley & Sons.
[4] G. J. Lieberman, D. B. Owen. *Tables of the Hypergeometric Probability Distribution*, 1961; Stanford University Press.
[5] J. K. Patel, Prediction Intervals-A Review, *Communications in Statistics, Theory and Methods*, vol 18, 1989, pp 2393–2465.
[6] J. K. Patel, V. A. Samaranayake, Prediction Intervals for Some Discrete Distributions, *J. Quality Technology*, vol 23, 1991, pp 270–278.
[7] S. S. Shapiro, H. Zahedi, Bernoulli Trials and Discrete Distributions, *J. Quality Technology*, vol 22, 1990, pp 173–253.

AUTHORS

Dr. Lee J. Bain; Department of Mathematics and Statistics; University of Missouri; Rolla, Missouri 65401-0249 USA.

Lee J. Bain is a Professor of Mathematics and Statistics at the University of Missouri-Rolla. He has published numerous articles in reliability and life testing, and is the author of *Statistical Analysis of Reliability and Lifetesting Models*, Marcel Dekker 1978; and co-author of *Introduction to Probability and Mathematical Statistics*, Duxbury, 1987.

Dr. Jagdish K. Patel; Department of Mathematics and Statistics; University of Missouri-Rolla, Missouri 65401-0249 USA.

Jagdish K. Patel is Professor of Statistics. He got his PhD from University of Minnesota in 1968. His research interests are in ranking and selection problems and statistical reliability. He is a co-author of *Handbook of Statistical Distributions*, Marcel Dekker (1976) and *Handbook of the Normal Distribution*, Marcel Dekker (1982).

Manuscript TR92-044 received 1992 March 9; revised 1992 August 10.

IEEE Log Number 04635

◄TR►