A LARGER SAMPLE SIZE IS NOT ALWAYS BETTER!!!

Nagaraj K. Neerchal
Department of Mathematics and Statistics
University of Maryland Baltimore County, Baltimore, MD 21250

Herbert Lacayo and Barry D. Nussbaum
United States Environmental Protection Agency
Washington, DC 20460

ABSTRACT

In a previous paper Neerchal, Lacayo and Nussbaum (2007) explored the behavior of the well-known problem of finding the optimal sample size for obtaining a confidence interval of a pre-assigned precision (or length) for the proportion parameter of a finite or infinite binary population. We illustrated some special problems that arise due to the discreteness of the population distribution and precision that is measured by the length of the interval rather than by the variance. Specifically, the confidence level of an interval of fixed length does not necessarily increase as the sample size increases. However, when such confidence levels are computed using normal approximations, we see a monotonic behavior. In this paper, we consider the corresponding problem under the Poisson approximation and show that for this distribution monotonicity does not hold and one should be beware of this seeming peculiarity in recommending sample sizes for studies involving estimation of means or proportions.

Keywords and Phrases: Confidence intervals, Poisson distribution, Binomial distribution, Hypergeometric distribution, Optimal sample size

1. INTRODUCTION

In many sample situations, the parent population size is relatively small, (say $N < 100$). For example, the United States Environmental Protection Agency (USEPA) routinely audits certain small databases. Further, if sampling is expensive, a customer may request the smallest sample size that attains or exceeds a specified confidence interval (CI) where that CI has a specified precision denoted by "τ", or d. [A more formal statement of this problem a la Cochran (1975) will follow later.] This question as stated by one customer is as follows: "How large a sample do I need to estimate the error rate of a specific data base." This is a "fairly straight forward" simple random sampling without replacement (SRSWOR, i.e. the Hypergeometric Distribution) problem. It is considered by us in a previous paper namely Neerchal, Lacayo and Nussbaum (2007).

In that paper, we found, contrary to our expectations, that increasing the sample size did not always increase the magnitude of the confidence level. For example as shown in Table 1., of the Appendix, for a population of N=90 and a desired precision of .02 (i.e. CI=.04), we see that the confidence level is **NOT** monotone, but rather goes up and down as the sample size increases.

This unexpected non-monotonic up-down-up behavior is observed for the Hypergeometric and Binomial Distributions, both discreet. On the other hand, the confidence level for the normal distribution, which is often used to approximate binomial probabilities, is monotonically increasing. In this paper we will investigate the monotonicity (or the lack thereof) of Poisson distribution.

2. A FORMAL STATEMENT OF THE PROBLEM

We consider the statistical inference problem of obtaining an interval estimate of a pre-assigned length (also referred to as precision) for the mean of a population. The objective is to provide the optimal sample size that will achieve a desired confidence level. A preliminary estimate of μ, the population mean, is available a priori. Suppose $\overline{X}_n$ denotes the sample mean from a sample of size n; then, we are looking for the smallest n such that

$$P\left(\mid \overline{X}_n - \mu \mid < \tau\right) \geq 1 - \alpha. \qquad (1)$$

Thus, the confidence interval is of fixed length 2τ around the sample mean and has at least $100(1-\alpha)\%$ confidence level.

If the population is finite (size N) and the sample is obtained by simple random sampling without replacement, (SRSWOR), the confidence level (1) above is given by summing up the appropriate terms from a hypergeometric distribution. That is.

$$P\left(\mid \overline{X}_n - \mu \mid < \tau\right) = \sum_{j=\lfloor np - n\tau \rfloor + 1}^{\lceil np + n\tau \rceil - 1} \frac{\binom{[Np]}{j}\binom{N - [Np])}{n-j}}{\binom{N}{n}} \qquad (2)$$

where n denotes the sample size, and where the ceiling and floor functions [$\lceil\ \rceil$,$\lfloor\ \rfloor$] indicate the smallest integer less than or equal to the quantity inside the brackets for the ceiling function and similarly the floor function indicates the largest integer not less than the quantity in the floor brackets. Suppose we assume that either the population is infinite or the sampling is done with replacement; then we can use the binomial distribution to compute the confidence level given in (1). That is,

$$P\left(\mid \overline{X}_n - \mu \mid < \tau\right) = \sum_{j=\lfloor n\lambda - n\tau \rfloor + 1}^{\lceil n\lambda + n\tau \rceil - 1} \binom{n}{j} p^j (1-p)^{n-j} \qquad (3)$$

Where ($\lceil\ \rceil$,$\lfloor\ \rfloor$) in the upper and lower limits of the summation symbol above denote ceiling and floor functions respectively. Of course, for large n, it is also common to use Normal approximation. Even elementary textbooks meant for the first course in Statistics contain elaborate descriptions of "Normal approximation to binomial distribution with or without correction". That is,

$$P\left(\mid \overline{X}_n - \mu \mid < \tau\right) \cong P\left(\frac{-n\tau}{\sqrt{np(1-p)}} < Z < \frac{n\tau}{\sqrt{np(1-p)}}\right) \qquad (4)$$

where Z denotes a standard normal random variable. The advantage of normal approximation is that, we can obtain an explicit formula for the optimal sample size to achieve the desired confidence level. As shown in Cochran (1977),

$$n_{opt} = \frac{z^2{}_{\alpha/2}\, p(1-p)}{\tau^2}$$

In Neerchal, Lacayo and Nussbaum(2007), we show that the normal approximation formula (4) for the confidence level is monotonically increasing as the sample size increases, while the exact formulas (2) and (3) correspond to an up-and-down (a saw tooth shape) growth pattern. Consequently, one needs to use caution when rounding up sample size formulas.

In this paper, we consider the same inference problem under the Poisson distribution, another popular distribution used widely in practical applications. The Poisson distribution is also used to approximate binomial when the sample size is large and the probability of success is small. We let $X_1, X_2 \cdots, X_n$ denote a simple random sample from a Poisson distribution with parameter λ, and consider $\left(\overline{X}_n - \tau, \overline{X}_n + \tau\right)$ as the fixed length confidence interval for λ. The confidence level of this interval is given by

$$P\left(\overline{X}_n - \tau < \lambda < \overline{X}_n + \tau\right) = P\left(\sum_{i=1}^{n} X_i - n\tau < n\lambda < \sum_{i=1}^{n} X_i + n\tau\right)$$

$$= P\left(n\lambda - n\tau < \sum_{i=1}^{n} X_i < n\lambda + n\tau\right)$$

$$= \sum_{i=\lfloor n\lambda - n\tau \rfloor + 1}^{\lceil n\lambda + n\tau \rceil - 1} e^{-n\lambda} \frac{(n\lambda)^i}{i!} \qquad (5)$$

where, once again, $\lceil\ \rceil, \lfloor\ \rfloor$ denote ceiling and floor functions as in (3).

We have also observed a similar behavior for the Poisson distribution as we did for the binomial distribution. In other words, the confidence level given in equation (5) is non-monotonic and its graph will have a saw tooth pattern. That is, if we let

$$\Delta P_n = P\left(\overline{X}_{n+1} - \tau < \lambda < \overline{X}_{n+1} + \tau\right) - P\left(\overline{X}_n - \tau < \lambda < \overline{X}_n + \tau\right),$$

then ΔP_n can actually be positive or negative as the sample size increases. This can be seen in Figures 1 through 3 of the Appendix.

3. RESULTS AND DISCUSSION

It is straightforward to compute the expression given in (5) using any software package which computes Poisson probabilities that provide plots of the relationships between confidence levels and sample size, for different combinations of λ and τ. [See Figures 1 to 3] The saw tooth pattern is obvious. This has

major consequences in determining recommended sample size. The usual practice of rounding up the optimal sample size formula to a higher integer may lead to a lower confidence level than desired.

The main thrust of the author's work is from the vantage point of applications, which focuses on the determination of optimal sample size. When the samples are expensive to obtain, as it was in the motivating example of US-EPA's auditing case study mentioned in the introduction, it would be quite costly if the additional samples taken actually drive down the confidence.

This would be like paying more and getting less!!

This preliminary investigation of Poisson distribution and our previous work leads to interesting research questions. We list some of them below.

1. Is this peculiar behavior of the confidence level of the fixed length confidence intervals true for all the common discrete distributions and false for all continuous distributions?
2. In our work so far, we focused on the fixed length confidence intervals and corresponding optimal "sample size determination problem" a la Cochran (1977). Another commonly used approach is based on looking at the coverage probabilities by specifying the Type I and Type II errors. In fact, for some of the commonly used discrete distributions, so-called exact confidence intervals are also available. See for example, page 247 (for Binomial) and page 251 (for Poisson) of Millard and Neerchal (2001). An interesting research question would be to ask "Would we see the saw tooth pattern in the confidence levels as a function the sample size for such confidence intervals as well?"

REFERENCES

Abramowitz and Stegun (1965). Handbook of Mathematical Functions. Dover Publications, Inc. New York.

Johnson, N.L. and Kotz, S. (1969). Distributions in Statistics: Discrete Distributions. Houghton Mifflin Company, New York.

Cochran, W. G. (1977). Sampling Techniques, 3rd ed., Wiley, New York.

Brown, L. D, Cai T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, Vol. 16, No. 2, 101-133.

Millard. S. M. and Neerchal, N. K. (2001). Environmental Statistics with S-PLUS. CRC/Chapman Hall, Boca Raton, FL.

Neerchal, N. K., Lacayo, H. and Nussbaum, B. D. (2007). Is a Large Sample Size Always Better? American Journal of Mathematics and Management Sciences. (In process).

APPENDIX: TABLES AND GRAPHS

TABLE 1. Exact Confidence Levels (Probability that the mean will be in the confidence interval specified by the precision tau) for shortest intervals around the sample mean, and the sample sizes proposed by various commercial programs for precision 10%

	Exact Confidence Level for indicated sample size n and precision [i.e. tau]			
Confidence Interval [i.e. 2*tau=2precision]	**n= 47**	**n = 60**	**n = 63**	**n=67**
0.02	0.7233	0.8799	0.8314	0.7457
0.04	0.9047	0.8799	0.9698	0.9454
0.06	0.9047	0.9811	0.9698	1.0000

Figure 1 Plot of confidence level [$P(|\bar{X}_n - \mu| < \tau)$] vs Sample Size n.
Length of CI: 2*tau , tau=.1, and Lambda= 1

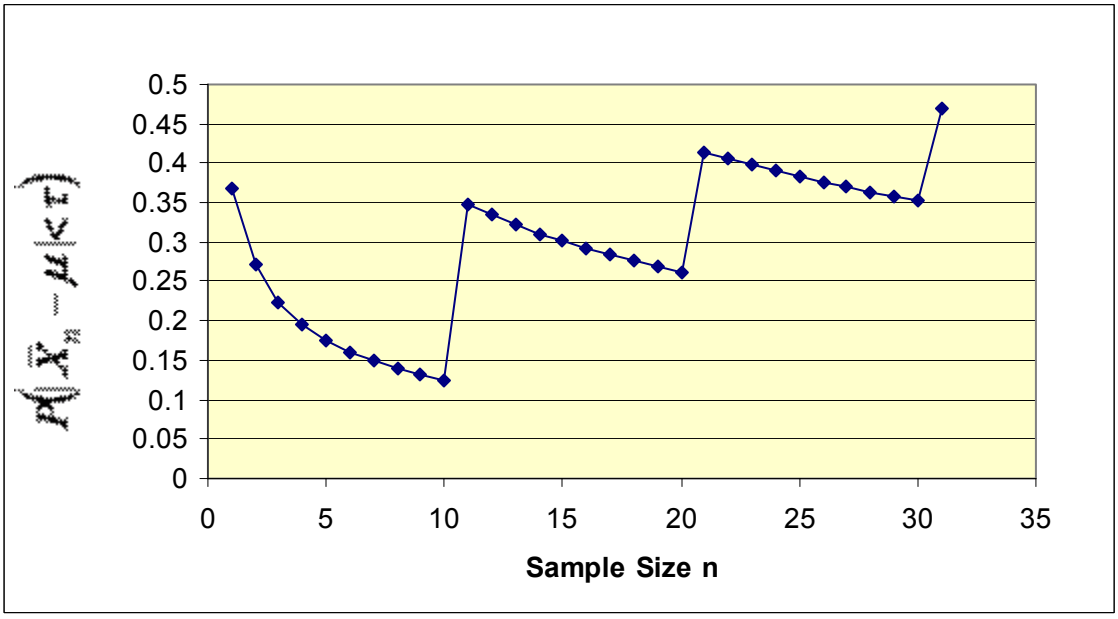


Table 2 Some raw data that may helps explain "jagadness" in Fig. 1

tau	lambda	n	up	lo	nlambda	prob1	prob2	n	conf
0.1	1	1	1	0	1	0.735759	0.367879	1	0.367879
0.1	1	2	2	1	2	0.676676	0.406006	2	0.270671
0.1	1	3	3	2	3	0.647232	0.42319	3	0.224042
0.1	1	4	4	3	4	0.628837	0.43347	4	0.195367
0.1	1	5	5	4	5	0.615961	0.440493	5	0.175467
0.1	1	6	6	5	6	0.606303	0.44568	6	0.160623
0.1	1	7	7	6	7	0.598714	0.449711	7	0.149003
0.1	1	8	8	7	8	0.592547	0.452961	8	0.139587
0.1	1	9	9	8	9	0.587408	0.455653	9	0.131756
0.1	1	10	10	9	10	0.58304	0.45793	10	0.12511
0.1	1	11	12	9	11	0.688697	0.340511	11	0.348186
0.1	1	12	13	10	12	0.681536	0.347229	12	0.334306
0.1	1	13	14	11	13	0.675132	0.353165	13	0.321967
0.1	1	14	15	12	14	0.66936	0.358458	14	0.310902
0.1	1	15	16	13	15	0.664123	0.363218	15	0.300905
0.1	1	16	17	14	16	0.659344	0.367527	16	0.291816
0.1	1	17	18	15	17	0.654958	0.371454	17	0.283505
0.1	1	18	19	16	18	0.650916	0.37505	18	0.275866
0.1	1	19	20	17	19	0.647174	0.378361	19	0.268813
0.1	1	20	21	18	20	0.643698	0.381422	20	0.262276
0.1	1	21	23	18	21	0.716029	0.30168	21	0.414349

Figure 2 Plot of confidence level [$P(|\bar{X}_n - \mu| < \tau)$] vs Sample Size n.
Length of CI: 2*tau, tau=.1, and Lambda= .1

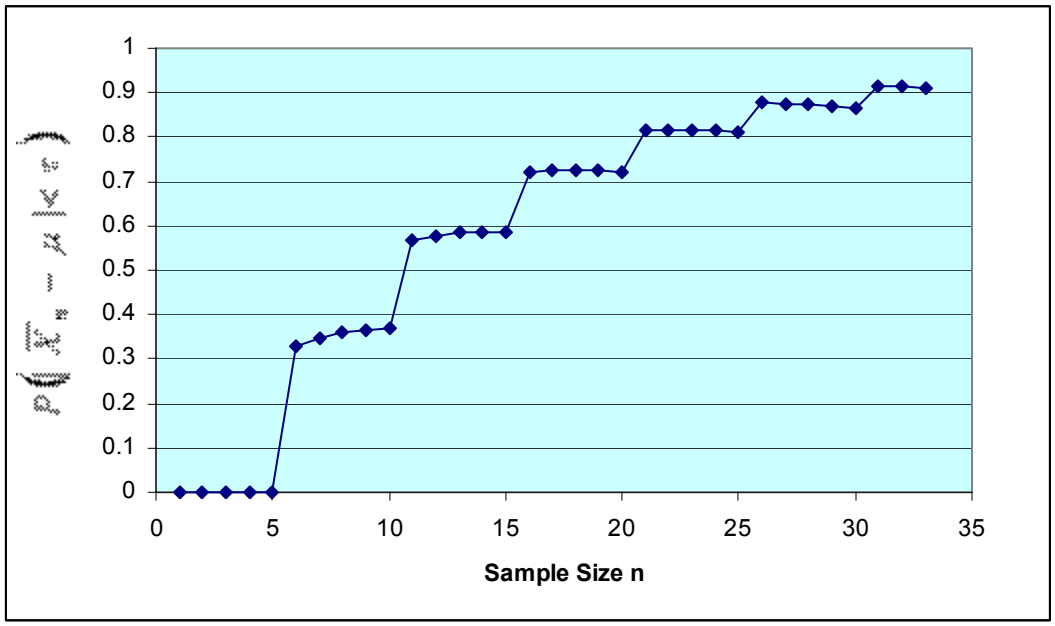
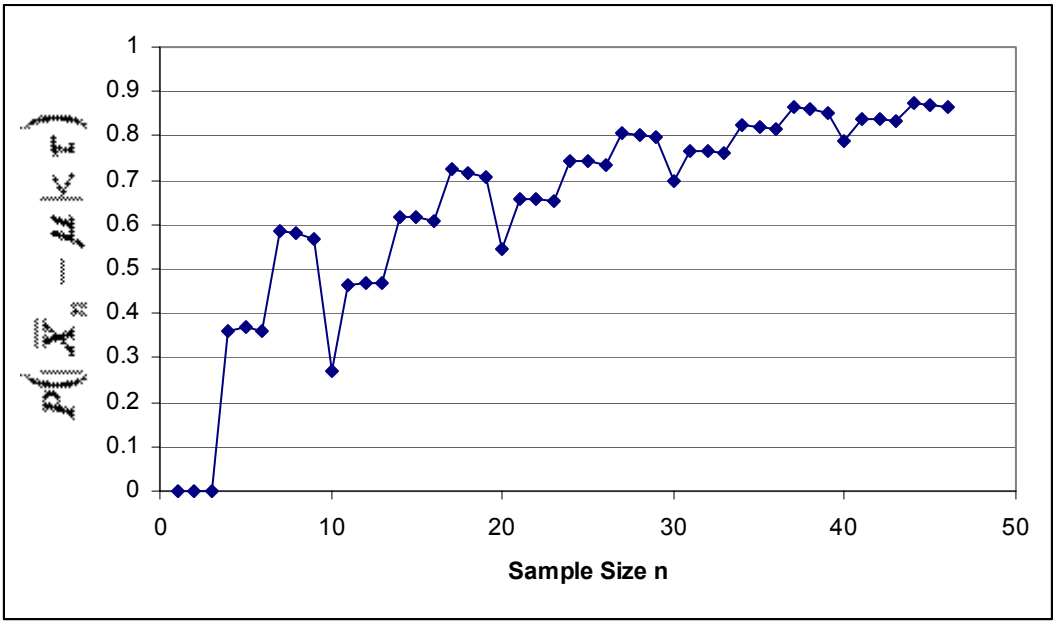


Table 3 Some raw data that may helps explain "jagadness" in Fig. 2

tau	lambda	up	lo	nlambda	prob1	prob2	n	conf
0.1	0.1	0	0	0.1	0.9048374	0.9048374	1	0
0.1	0.1	0	0	0.2	0.8187308	0.8187308	2	0
0.1	0.1	0	0	0.3	0.7408182	0.7408182	3	0
0.1	0.1	0	0	0.4	0.67032	0.67032	4	0
0.1	0.1	0	0	0.5	0.6065307	0.6065307	5	0
0.1	0.1	1	0	0.6	0.8780986	0.5488116	6	0.329287
0.1	0.1	1	0	0.7	0.844195	0.4965853	7	0.3476097
0.1	0.1	1	0	0.8	0.8087921	0.449329	8	0.3594632
0.1	0.1	1	0	0.9	0.7724824	0.4065697	9	0.3659127
0.1	0.1	1	0	1	0.7357589	0.3678794	10	0.3678794
0.1	0.1	2	0	1.1	0.9004163	0.3328711	11	0.5675452
0.1	0.1	2	0	1.2	0.8794871	0.3011942	12	0.5782929
0.1	0.1	2	0	1.3	0.8571125	0.2725318	13	0.5845807
0.1	0.1	2	0	1.4	0.8334977	0.246597	14	0.5869008
0.1	0.1	2	0	1.5	0.8088468	0.2231302	15	0.5857167
0.1	0.1	3	0	1.6	0.9211865	0.2018965	16	0.71929
0.1	0.1	3	0	1.7	0.9068106	0.1826835	17	0.724127

Figure 3. Plot of confidence level [$P(|\bar{X}_n - \mu| < \tau)$] vs Sample Size n.
Length of CI: 2*tau, tau=.1, and Lambda=.2



Discussion of Figures 1 and 2

In Figure 1, we note a peculiar pattern as the sample size increases. For instance, starting with a sample size of 11, we see [from Table 2] that the confidence level is .348… As the sample size increases to 12, 13, …20, the confidence level actually decreases. However, when the sample size goes from 20 to 21, there is a marked increase in the confidence level [from .262… to .414…]. This pattern appears to repeat in cycles of 10.

In addition, in Figure 2, we note a similar peculiar pattern as the sample size increases. For instance, starting with a sample size of 6, we see from Table 3 that the confidence level is .329…… As the sample size increases to 7, 8, 9, 10, the confidence level increases. However, when the sample size goes from 10 to 11, there is a marked increase in the confidence level [from .367… to .567…]. This pattern appears to repeat in cycles of 5.