The Sample Size for Estimating the Binomial Parameter with a Given Margin of Error

By B.K. Ghosh and Wei-Min Huang

Lehigh University

Consider the problem of constructing an estimator $T(X_n)$ for the binomial p and determining the smallest sample size N_T such that, for specified values of (γ, δ) and for *all* $n \geq N_T$, the interval $T(X_n) \pm \delta$ contains the true value of p with a minimum probability γ. In principle, any conventional $100\gamma\%$ confidence interval for p can be adapted to solve the present problem. We show that, for small δ, all known confidence intervals effectively lead to the estimator $T_b(X_n) = (X_n + \frac{1}{2}b)(n+b)^{-1}$ with the sample size $N_b \approx (z/2\delta)^2 + \delta^{-1}$, where z is the $\frac{1}{2}(1+\gamma)$ quantile of the standard normal distribution and $b \geq 0$ is a specified constant which does not depend on n. The major purpose of this paper is to propose the estimator $T^*(X_n) = (X_n + \frac{1}{2}a_\gamma\sqrt{n})(n + a_\gamma\sqrt{n})^{-1}$ with sample size $N^* \approx \{(z/2\delta) - a_\gamma\}^2 + \delta^{-1}$, where $a_\gamma = 1$ when $\gamma \geq .917$ and $0 \leq a_\gamma < 1$ is defined in (1.10) when $\gamma < .917$. The proposed method is more efficient in the sense that $N^* < N_b$ for any $b \geq 0$. These asymptotic conclusions are shown to be quite adequate for arbitrary values of $\delta \in [.01, .10]$ and $\gamma \in [.90, .99]$ by making exact calculations for the minimum coverage probabilities of $T_b(X_n) \pm \delta$ and $T^*(X_n) \pm \delta$ as well as for N_b and N^*.

1. Introduction. An ubiquitous but rarely researched problem of practical statistics arises in the context of a binomial distribution. Suppose we can observe the number of "successes" X_n in n independent "trials", each trial having the same unknown probability of success p, $0 < p < 1$. Let $\gamma \in (0,1)$ be a specified *confidence level* and $\delta \in (0, \frac{1}{2})$ be a specified *margin of error*. Then the problem is to construct an *estimator* $T(X_n)$ for p and predetermine the (smallest) *sample size* N_T such that

$$(1.1) \qquad \inf_{0<p<1} \gamma_T(p,n) \geq \gamma \quad \text{for } \textit{all} \quad n \geq N_T,$$

where

$$(1.2) \qquad \gamma_T(p,n) = P_p(|T(X_n) - p| \leq \delta)$$

is the *coverage probability* of the interval $T(X_n) \pm \delta$. The implication of (1.1) is, of course, that $T(X_n) \pm \delta$ contains the true value of p with a minimum probability γ for any $n \geq N_T$. We allow $T(X_n)$ to involve γ and, clearly, $\gamma_T(p,n)$ will involve δ while N_T will generally depend on both γ and δ. We will sometimes qualify $\gamma_T(p,n)$ and N_T with the word *exact* in order to emphasize that (1.2) and N_T are computed under the *binomial* distribution of X_n. The problem just described is encountered almost daily in opinion polls, market research, clinical trials and quality control, where one usually chooses $.90 \leq \gamma \leq .99$ and $.01 \leq \delta \leq .10$. Note that for any given $\delta \geq \frac{1}{2}$ the trivial choice $T(X_n) = \frac{1}{2}$ with $N_T = 0$ provides a solution. Clearly, if $T(X_n)$ and $T^*(X_n)$ are two competing estimators and their sample sizes satisfy $N_T > N_{T^*}$, one should prefer to use $T^*(X_n) \pm \delta$. Indeed, a main