QUESTION ABOUT CHEYBSHEV'S INEQUALITY AND THE CENTRAL LIMIT THEOREM

FRED J. HICKERNELL

Suppose one wishes to approximate the mean, $\mu = E(X)$, by the sample average of i.i.d. draws, $X_1, X_2, \ldots$, i.e.,

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1}$$

An important question that springs to mind is how to choose the right sample size, n, to achieve a specified tolerance, ϵ. The random estimator, $\hat{\mu}_n$ has mean μ and variance

$$\operatorname{var}(\hat{\mu}_n) = \frac{\sigma^2}{n},$$

where $\sigma^2 = \operatorname{var}(X) = \operatorname{var}(X_i)$.

In practice, one often invokes the Central Limit Theorem to determine sample size. Given a significance level or uncertainty tolerance, α, one has

$$\text{Prob}\left[|\hat{\mu}_n - \mu| \le \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right] \approx 1 - \alpha.$$

Thus one has the *approximate* probabilistic result

$$\text{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \approx 1 - \alpha \qquad \text{for } n = N_N(\epsilon, \alpha, \sigma^2) := \left\lceil \left(\frac{z_{\alpha/2}\sigma}{\epsilon}\right)^2 \right\rceil. \tag{2}$$

The above is exact if the X_i are i.i.d. normal, however, in general this result is only approximate and relies on the unknown σ. The above observations are formalized in the proposition below.

The first of these drawbacks may be removed by turning to Chebyshev's inequality. Using Chebyshev's inequality yields an exact upper bound rather than approximate one:

$$\text{Prob}\left[|\hat{\mu}_n - \mu| < \frac{\sigma}{\sqrt{n\alpha}}\right] \ge 1 - \alpha.$$

Thus, a proper choice of sample size guarantees that the estimate is within the tolerance of the true answer with probability $1 - \alpha$:

$$\text{Prob}\left[|\hat{\mu}_n - \mu| < \epsilon\right] \ge 1 - \alpha \qquad \text{for } n = N_C(\epsilon, \alpha, \sigma^2) := \left\lceil \frac{\sigma^2}{\alpha\epsilon^2} \right\rceil. \tag{3}$$

This sample size is typically much larger than N_N, since $1/\sqrt{\alpha}$ is typically much larger than $z_{\alpha/2}$.

Question 1: Is there a tighter inequality than Chebyshev's, when the random variable is known to be a sum of i.i.d. random variables, but not assuming boundedness (like Hoeffding's inequality), only the existence of the first two or more moments? From Ghosh and Meeden (1977) we know that it is not tight.

Is there something that tells us now close we are to the Central Limit Theorem as a function of n or the moments?

Although the Chebyshev result is exact for general distributions, it still depends on the typically unknown σ^2. This is why in practice one normally uses observed function values observed to approximate the σ^2 by the sample variance, as follows:

$$\hat{v}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2. \tag{4}$$

This means that we now have an *adaptive* algorithm. One might choose an initial sample of size n_0, and use it to estimate σ^2 by $\hat{v}_{n_0}$. Then one chooses an *independent* sample of size $n = N_C(\epsilon, \alpha, \hat{v}_{n_0})$ or $n = N_N(\epsilon, \alpha, \hat{v}_{n_0})$ to compute $\hat{\mu}_n$ the final estimate of μ.

Unfortunately, once we approximate σ^2 by $\hat{v}_n$, we again have inexact results. However, they can be made exact by using Cantelli's inequality and the variance of $\hat{v}_n$:

$$\begin{aligned} 1-\alpha &\le \text{Prob}\left[\hat{v}_n - \sigma^2 \ge -\sigma^2\sqrt{\frac{1}{n}\left(\kappa + \frac{2n}{n-1}\right)\left(\frac{1-\alpha}{\alpha}\right)}\right] \\ &= \text{Prob}\left[\frac{\hat{v}_n}{1-\sqrt{\left(\kappa+\frac{2n}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}} \ge \sigma^2\right] \end{aligned}$$

Thus,

$$\text{Prob}\left[\hat{\sigma}_{\text{up}}(n,\alpha,\kappa) \ge \sigma\right] \ge 1-\alpha, \qquad \text{where } \hat{\sigma}^2_{\text{up}}(n,\alpha,\kappa) = \frac{\hat{v}_n}{1-\sqrt{\left(\kappa+\frac{2n}{n-1}\right)\left(\frac{1-\alpha}{n\alpha}\right)}}, \tag{5}$$

provided that

$$\begin{aligned} 1 &> \left(\kappa + \frac{2n}{n-1}\right)\left(\frac{1-\alpha}{n\alpha}\right) \\ \frac{n\alpha}{1-\alpha} &> \kappa + \frac{2n}{n-1} \\ \kappa &< \frac{n\alpha}{1-\alpha} - \frac{2n}{n-1} =: \kappa_{\text{poss}}. \end{aligned}$$

Theorem 1. *For a given positive constant, $\kappa_{\max}$, define the set of random variables with finite fourth moments:*

$$\mathcal{Y} = \{X : \text{kurt}(X) = \kappa \le \kappa_{\max},\},$$

where the kurtosis of the function is defined in (6)*. Suppose that one has an error tolerance, ϵ, and an uncertainty tolerance, α. Let $\alpha_1 = 1 - \sqrt{1-\alpha}$. Pick any $n_0 > 1$ satisfying*

$$\kappa_{\max} < \frac{n_0\alpha_1}{1-\alpha_1} - \frac{2n_0}{n_0-1},$$

and compute the sample variance, $\hat{v}_{n_0}$ of a simple random sample of size n_0. Use this to compute $\hat{\sigma}^2_{\text{up}}(n_0, \alpha_1, \kappa_{\max})$ by (5)*. Next choose an independent random sample of size n and compute $\hat{\mu}_n$, the simple Monte Carlo estimator of μ. A probabilistic error bound is given by*

$$\text{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] = 1-\alpha \qquad \textit{for } n \ge N_C(\epsilon, \alpha_1, \hat{\sigma}^2_{\text{up}}(n_0, \alpha_1, \kappa_{\max})).$$

Proof. By (5) it follows that $\hat{\sigma}_{\text{up}}(n,\alpha,\kappa) \ge \sigma$ with probability $1-\alpha_1$.

□

The point is that now we have a guaranteed (with a certain confidence) program for getting an approximation to the mean if we know that our random variable does not have too large a kurtosis. But at least we do not have to know the varaiance in advance.

Question 2: Is there a tighter bound for the sample variance than Cantelli's Theorem?

Question 3: Is this all known and written down somewhere? My real aim is to apply this kind of analysis to problems where we use something more sophisticated as our estimator.

Appendix

Theorem 2 (Chebyshev's and Cantelli's Inequalities)**.** *Let Z be any random variable with mean μ_Z and variance σ_Z^2. Then for all $\alpha > 0$, Chebyshev's inequality states that*

$$\text{Prob}\left[|Z-\mu_Z| \geq \frac{\sigma_Z}{\sqrt{\alpha}}\right] \leq \alpha, \qquad \text{Prob}\left[|Z-\mu_Z| < \frac{\sigma_Z}{\sqrt{\alpha}}\right] \geq 1-\alpha.$$

Cantelli's inequality states that

$$\text{Prob}\left[Z-\mu_Z \geq \sigma_Z\sqrt{\frac{1-\alpha}{\alpha}}\right] \leq \alpha, \qquad \text{Prob}\left[Z-\mu_Z < \sigma_Z\sqrt{\frac{1-\alpha}{\alpha}}\right] \geq 1-\alpha,$$

$$\text{Prob}\left[Z-\mu_Z \leq -\sigma_Z\sqrt{\frac{1-\alpha}{\alpha}}\right] \leq \alpha, \qquad \text{Prob}\left[Z-\mu_Z > -\sigma_Z\sqrt{\frac{1-\alpha}{\alpha}}\right] \geq 1-\alpha.$$

Theorem 3. *Let $\hat{v}_n$ be the sample variance as defined in* (4)*. It's variance is*

$$\text{var}(\hat{v}_n^2) = \frac{\sigma^4}{n}\left(\kappa + \frac{2n}{n-1}\right),$$

where κ denotes the kurtosis*:*

$$\kappa := \text{kurt}(X) := \frac{\gamma}{\sigma^4} - 3, \qquad \gamma := E[(X-\mu)^4]. \tag{6}$$

References

Ghosh M, Meeden G (1977) On the non-attainability of chebyshev bounds. Amer Statist 31:35–36