COMMUN. STATIST.—SIMULA., 28(1), 177–188 (1999)

ACCURATE APPROXIMATION TO THE EXTREME ORDER STATISTICS OF GAUSSIAN SAMPLES

Chien-Chung Chen & Christopher W. Tyler

Smith-Kettlewell Eye Research Institute
San Francisco, CA 94115

Key Words: gamma approximation; expected value; probability density function; maximum; minimum;

ABSTRACT

Evaluation of the integral properties of Gaussian Statistics is problematic because the Gaussian function is not analytically integrable. We show that the expected value of the greatest order statistics in Gaussian samples (the max distribution) can be accurately approximated by the expression $\Phi^{-1}(0.5264^{1/n})$, where n is the sample size and Φ^{-1} is the inverse of the Gaussian cumulative distribution function. The expected value of the least order statistics in Gaussian samples (the min distribution) is correspondingly approximated by $-\Phi^{-1}(0.5264^{1/n})$. The standard deviation of both extreme order distributions can be approximated by the expression $0.5[\Phi^{-1}(0.8832^{1/n}) - \Phi^{-1}(0.2142^{1/n})]$. We also show that the probability density function of the extreme order distribution can be well approximated by gamma distributions with appropriate parameters. These approximations are accurate, computationally efficient, and readily implemented by build-in functions in many commercial mathematical software packages such as MATLAB, Mathematica, and Excel.

Copyright © 1999 by Marcel Dekker, Inc. www.dekker.com

INTRODUCTION

Consider n samples x_1, x_2, ...,x_n from a standard Gaussian distribution, N(0,1). The extreme order distributions are the distributions of the greatest and the least values among n samples from the Gaussian distribution. Let $x_{max} = \max_i(x_i)$, i = 1,2,...,n be the greatest of the n sample values. The probability distribution of x_{max} has the density function

$$PDF(x_{max}) = n\, \Phi(x_{max})^{(n-1)}\, \phi(x_{max}) \tag{1}$$

where $\phi(x)$ is the probability distribution function (PDF) and $\Phi(x)$ is the cumulative distribution function (CDF) of the standard Gaussian distribution (Bain & Engelhardt, 1987). The greatest order distribution PDFs of selected sample sizes are shown in Figure 1.

For the least of the n samples $x_{min} = \min_i(x_i)$, i = 1,2,...,n, has the probability density distribution

$$\begin{aligned} PDF(x_{min}) &= n\, [1-\Phi(x_{min})]^{(n-1)}\, \phi(x_{min}) \\ &= n\, \Phi(-x_{min})^{(n-1)}\, \phi(-x_{min}). \end{aligned} \tag{2}$$

Extreme order distributions are widely used in fields such as biology, psychophysics, economics, seismology, signal processing and analysis of parallel distributed noisy systems. It is particularly relevant in the analysis of stochastic resonance phenomena, where the addition of noise can increase detectability of a signal derived from a nonlinear system (Bulsara et al., 1991; Bezrukov & Vodyanoy, 1997). However, since the Gaussian CDF $\Phi(x)$ cannot be expressed in terms of elementary functions, it is difficult to integrate $\Phi(x)$ analytically, and thus analytic solutions to the moments of the extreme order distribution are difficult to find. Statisticians have made efforts to find analytical solutions to expected value and standard deviation of the extreme order distribution with a recurrence method (Jones, 1948; Ruben, 1954; Bose & Gupta, 1959; David, 1963). Although this method is successful for small sample sizes, it is tedious and fails for sample size n >= 6 (Arnold & Balakrishnan, 1989; Harter & Balakrishnan, 1996), which makes it of limited utility.

The expected value and standard deviation of the extreme order distributions of Gaussian samples have been tabled for selective sample sizes by numerical

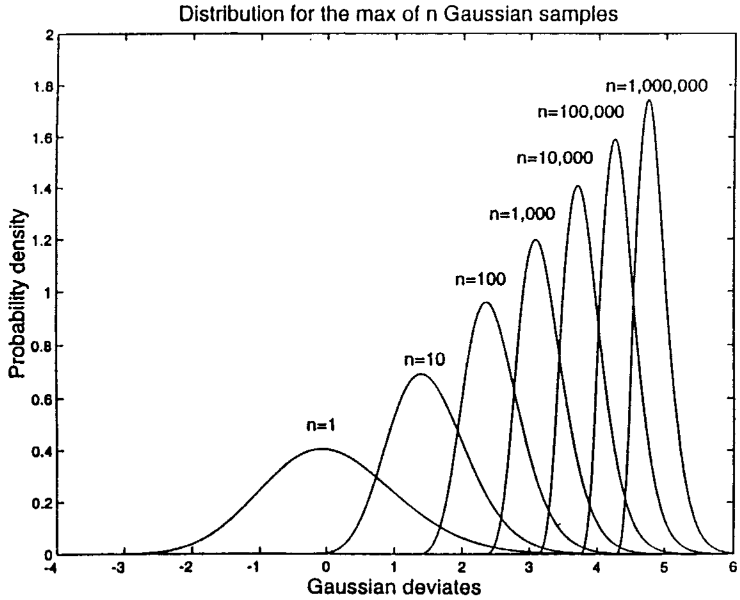

FIG. 1. Probability density functions for the greatest order values of Gaussian samples with sample sizes n from 1 to 1,000,000 in decade steps.

integration (Harter, 1961; Parrish, 1992a, b). Those tables are not very practical because they only list selected sample sizes and thus we still have no access to the expected value and the standard deviation of the extreme order distribution of arbitrary sample sizes. Moreover, the accuracy of numerical integration depends on the range of the independent variable and size of the bin used for integration. The wider the range and the smaller the bin size, the more accurate is the integration. To increase the range and decrease the bin size correspondingly increases the computation time. Thus, accurate numerical integration is quite time consuming.

Blom (1958) suggested an expression to approximate the expected value E_n of the greatest order distribution (max) numerically:

$$B_n = \Phi^{-1}\left(\frac{i-\alpha}{n-2\alpha+1}\right) \qquad (3)$$

However, the constant a changes continuously with the sample size n. Moreover, there is no simple relation between α and n. Thus, his method fails to compute the expected value of the max distribution with any arbitrary sample size.

For the parameter of the variance of the max distribution σ_n^2 Pelli (1985) suggested the approximation $\frac{\pi^2}{12\ln(n+1)}$ where n is the sample size. However, as we will show, Pelli's approximation has limited accuracy.

We were also disappointed that there was no good algorithm to compute and approximate the form of the PDFs of the extreme order distributions in the literature. Thus, it is impossible to calculate many statistics, such as the difference or probabilistic combination of two extreme order statistics of different sample sizes.

Our goal here is to find an accurate approximation formula for the expected value, standard deviation and PDF of the extreme order distributions in Gaussian samples. To make these approximations more practical, we develop expressions that can be computed with the built-in functions provided by many commercial mathematical or spread sheet software packages such as MATLAB, Mathematica, or Excel. We compare our estimation of the extreme order distributions to those obtained with standard numerical integration.

METHOD

EXPECTED VALUE

The PDF of the max distribution in eq. 1 can be rewritten as

$$PDF(x_{max}) = \frac{d}{dx_{max}}[\Phi(x_{max})]^n \tag{4}$$

Thus, the CDF of the max distribution is $\Phi(x_{max})^n$. At the median of the max distribution, $\tilde{x}_{max}$,

$$\Phi(\tilde{x}_{max})^n = 0.5 \tag{5}$$

From eq. 5, the solution for the median of the max distribution is

$$\tilde{x}_{max} = \Phi^{-1}(0.5^{1/n}) \tag{6}$$

where Φ^{-1} is the inverse function of the Gaussian CDF, Φ. If the max distribution were symmetrical about the expected value, eq. 6 should provide a good estimation of the expected value. However, since the max distribution is skewed to the left, a correction to eq. 6 is needed. We find that the expression

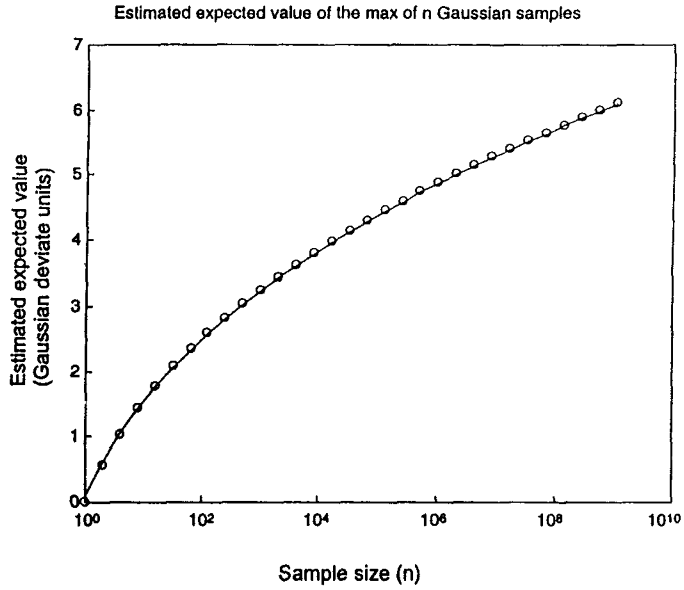

FIG. 2. The open circles denote the expected values of the max distribution estimated by the trapezoidal numerical integration method. The smooth curve is the expected value estimated by the expression $\Phi^{-1}((0.5264)^{1/n})$ (eq. 7). The horizontal axis is the sample size.

$$E_n = \Phi^{-1}((0.5+\varepsilon)^{1/n}),\ \varepsilon=0.0264 \qquad (7)$$

provides a good approximation to the expected value of the max distribution. The correction factor ε was obtained by a bisectional searching algorithm that optimizes the match of eq. 7 to the expected values. We may compare the expected value estimated from eq. 7 and the expected value estimated from numerical integration with trapezoidal rule (Press et al., 1986). The result is shown in Fig. 2 where the numerical integration is computed over the ranges from -6 to 10 standard deviations of the Gaussian in steps of 0.01. In Fig. 2, the open circles denote the expected value estimated from numerical integration over nine orders of magnitude of n and the curve represents the function of eq. 7 over sample size. The continuous curve in Fig. 3 shows the percentage difference between the estimation of eq. 7 and

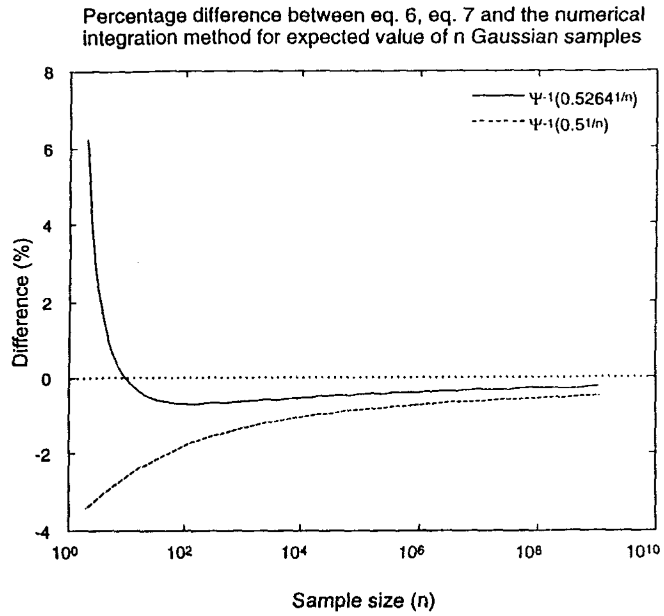

FIG. 3. The percentage difference between the expected values of the max distribution estimated by the trapezoidal numerical integration method and by eq. 6 and eq. 7. The continuous curve is the percentage difference between eq. 7 and the trapezoidal numerical integration method; the broken curve is the equivalent difference for eq. 6.

the numerical integration method while the broken curve shows the percentage difference between the sample median (eq. 6) and the numerical method. For the numerical integration estimation u_n and the eq. 7 approximation E_n, the percentage difference at sample size n is $(E_n-u_n)/u_n \times 100\%$. For eq. 7, the deviation is less than 2% for sample sizes $n > 5$, and less than 1% when $n > 10$. For $n < 5$, eq. 6 is more appropriate for the approximation, the deviation being confined to 3%. We note that the same approach may be elaborated to any desired degree of accuracy by expanding e in terms of a Taylor series of the variable n.

From eq. 2, it is clear that the expected value of the least order distribution (min) in Gaussian samples is just -1 times the expected value of the max distribution at the same sample size. Thus, $-E_n$ provides an approximation to the expected value of min distribution.

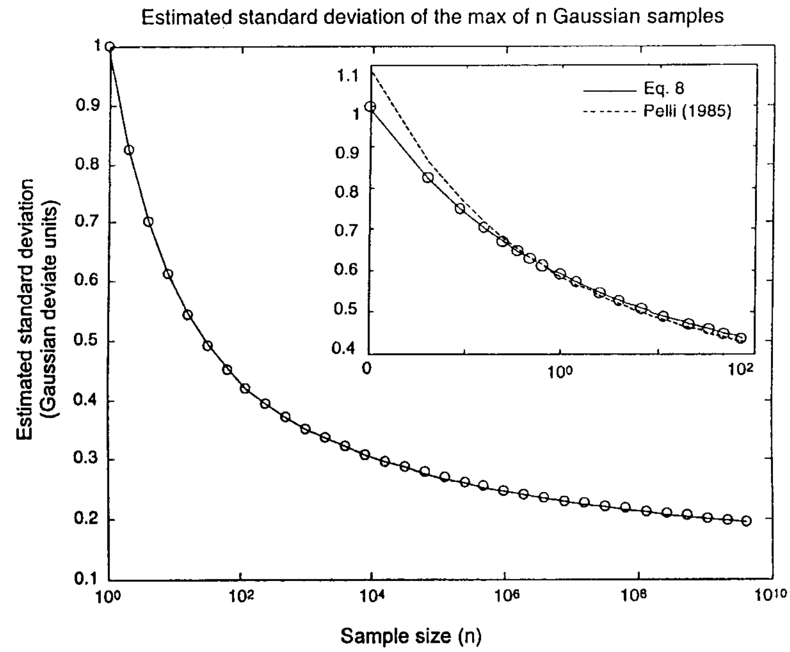

FIG. 4. The open circles denote the standard deviation of the max distribution estimated by the trapezoidal numerical integration method. The continuous curve is the expected value estimated by eq. 8. The horizontal axis is the sample size. The inset shows the standard deviation estimated for small sample sizes by the numerical integration method (open circles), eq. 8 (continuous curve) and Pelli's approximation (broken curve).

STANDARD DEVIATION

The variable of interest in applications such as signal detection is the standard deviation of the max distribution. If the max distribution were symmetrical, the range $0.5[\Phi^{-1}((0.5+a)^{1/n})-\Phi^{-1}((0.5-a)^{1/n})]$, where $a = 0.3625$, should provide a good approximation to standard deviation. Again, since the max distribution is positively skewed, a slight modification can improve the estimate. We find that

$$s_n=0.5\ [\Phi^{-1}((0.5+a_+)^{1/n})-\Phi^{-1}((0.5-a_-)^{1/n})] \tag{8}$$

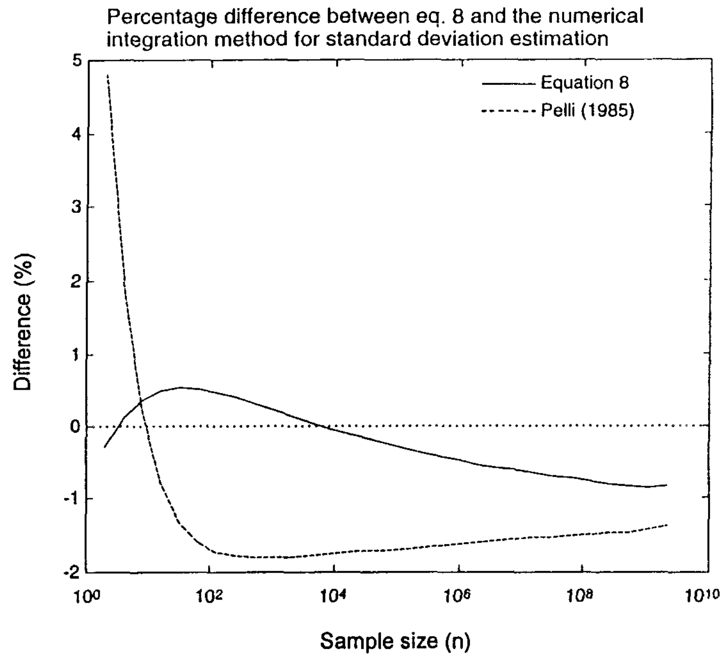

FIG. 5. The percentage difference between the standard deviation of the max distribution estimated by the trapezoidal numerical integration method, by eq. 8 and by Pelli's approximation.

where $a_+ = 0.3832$ and $a_- = 0.2858$, gives an excellent approximation to the standard deviation of the max distribution. We used a two dimensional searching algorithm to find the values for a_+ and a_- that minimize the least square error between the standard deviation estimated from eq. 8 and from numerical integration. Fig. 4 compares the standard deviation estimation from numerical integration (open circles) and from eq. 8 (smooth curve). Fig. 5 shows the difference between the two estimations. For the numerical integration estimation σ_n and the eq. 8 approximation s_n, the percentage difference at sample size n is $(s_n-\sigma_n)/\sigma_n \times 100\%$. For all sample sizes up to 1,000,000, the deviation is less than 0.5%. The standard deviation of the min distribution is the same as the standard deviation of the max distribution. We again note that the same approximation can be elaborated to any desired of level accuracy by expanding a_+ and a_- as Taylor series.

PROBABILITY DENSITY FUNCTION

For a more complete characterization of the extreme order statistics, we find that we can approximate the PDF of the max distribution by the PDF of the Gamma distribution:

$$g(x) = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{x-c}{\alpha}\right)^{\beta-1} e^{-(x-c)/\alpha};\ x > c;\ \alpha, \beta > 0. \tag{9}$$

where $\Gamma(\beta)$ is the gamma function with argument β. To approximate the max of n samples, we include the location parameter, c, given by the an empirical function, developed for this purpose to shift the PDF appropriately away from zero,

$$c = 2.8989 * \ln(\log_2(n)) - 4.4291. \tag{10}$$

It has been shown (e.g., Bain & Engelhardt, 1987) that the expected value μ and variance σ^2 of the gamma distribution with parameters c, α, and β are

$$\mu = \alpha\beta + c$$

$$\sigma^2 = \alpha^2\beta \tag{11}$$

Thus, combining the results of eq. 7, 8, 10 & 11, the max distribution of n Gaussian samples can be approximated by the gamma distribution with parameters

$$\alpha = \frac{s_n^2}{E_n - c}$$

$$\beta = \frac{(E_n - c)^2}{s_n^2} \tag{12}$$

where E_n is the estimated expected value from eq. 7 and s_n is the estimated standard deviation from eq. 8. The continuous curves in Fig. 6 show PDFs for the max distribution estimated by the gamma approximation and the dotted curves are PDFs computed by numerical integration. For the 8407 points in the 7 PDFs estimated, the mean square error is only 0.00033 and the Chi-square value 7.3401 with 8400 degree-of-freedom. Thus, this approximation is rather accurate.

DISCUSSION

Our methods provide simple and efficient means of approximating the expected value, the standard deviation and the complete PDF of the extreme order