2012 Fifth International Joint Conference on Computational Sciences and Optimization

Approximating the minimum distribution of two normally distributed variables each with the same mean and variance

Zhengbing He
School of Traffic and Transportation
Beijing Jiaotong University
Beijing, China
Email: he.zb@hotmail.com

Ailing Huang
School of Traffic and Transportation
Beijing Jiaotong University
Beijing, China
Email: alhuang@bjtu.edu.cn

***Abstract*—Several integrals impose excessive computational burden in the solution of the minimum distribution of two normally distributed variables each with the same mean and variance. To overcome the inefficiency, this paper first investigates the probability and maximum value of deviation occurrence between the normal distributions, and then proposes an approximation method of the mean and variance of the distribution. The test results show that the approximations give high accuracy in the range from 10 to 10000, and the more importance is that one can modify the fitting parameters in the method to obtain approximations for other ranges.**

***Keywords*-Approximation method; normally distributed variable; minimum distribution**

I. Introduction

The comparison of two normally distributed variables each with the same mean and variance is meaningful. The Poisson distribution is usually used to express discrete occurrences that take place during a time-interval of given length. For the Poisson distribution with a parameter greater than 10, the normal distribution with the same mean and variance that are equivalent to the parameter is a good approximation. The occurrences of discrete events can thus be approximately described using the normal distribution with the same mean and variance, and the comparison of two events can be derived from that of two normal distributions.

An approximation of the minimum of the two variables is useful. The numerical integral has to be applied repeatedly in the computational process without approximations, and the efficiency is low. A direct approximation to the expression is able to reduce the burden. It is critical to engineering sometimes.

Unfortunately, the relevant researches focusing on this approximation are few. Only when the variables follows the same normal distribution, the Gumbel distribution is considered as an asymptote of the minimum distribution of these variables[1][2]. To fill this void, this paper investigates the properties of the minimum distribution of two normally distributed variables each with the same mean and variance, and provides an approximation method of the mean and variance of the distribution. By using the approximation, the numerical integrals are avoided and the efficiency is greatly improved.

II. Exact solution

Let $X \sim \mathcal{N}(\mu_X, \mu_X)$ and $Y \sim \mathcal{N}(\mu_Y, \mu_Y)$ be the two normal distributions respectively with the same mean and variance, μ_X and μ_Y. The exact solution for the cumulative distribution function (CDF) of $Z = \min\{X, Y\}$ is

$$\begin{aligned} F_Z(z) &= 1 - \mathrm{P}\{\min(X,Y) > z\} \\ &= 1 - \mathrm{P}\{X > z, Y > z\} \\ &= 1 - [1 - F_X(z)][1 - F_Y(z)] \\ &= F_X(z) + F_Y(z) - F_X(z)F_Y(z), \end{aligned} \tag{1}$$

The corresponding probability density function (PDF) is

$$f_Z(z) = f_X(z) + f_Y(z) - f_X(z)F_Y(z) - f_Y(z)F_X(z), \tag{2}$$

The mean is

$$\mu_Z = \mu_X + \mu_Y - \int_{-\infty}^{\infty} z f_X(z) F_Y(z) dz - \int_{-\infty}^{\infty} z f_Y(z) F_X(z) dz, \tag{3}$$

The variance σ_Z^2 can be calculated by using the basic formula:

$$\sigma_Z^2 = \int_{-\infty}^{\infty} (z - \mu_Z)^2 f_Z(z) dz. \tag{4}$$

The term $F_X(z)F_Y(z)$ in Equation 1 is a normal product distribution. The approximation and numerical evaluation of the distribution has been presented in some works such as [3-5]. The mean and variance of the distribution, however, can not be obtained easily due to the complexity and limitation of results of these works. In addition, the equations of $F_Z(z)$ and $f_Z(z)$ can also be computed by some computing softwares without much computational burden (For example, the CDF and PDF of normal distributions can be computed directly in Matlab by using functions "normcdf" and "normpdf"). For μ_Z and σ_Z^2, however, no efficient can be utilized. One may have to do numerical integral. The computational

978-0-7695-4690-2/12 $26.00 © 2012 IEEE
DOI 10.1109/CSO.2012.31

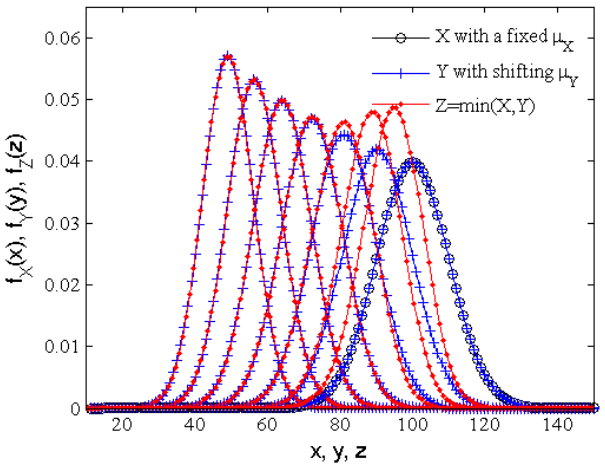


Figure 1. Distributions of Z corresponding to X with a fixed μ_X and Y with shifting μ_Y

efficiency is thus very low, especially for σ_Z^2 in which one numerical integral contains another. We will introduce an approximation of the mean and variance of Z in the paper.

III. Deviation occurrence and the maximum value

To have intuitive understanding, we demonstrate distributions of Z corresponding to X with a fixed μ_X and Y with shifting μ_Y in Figure 1. It can be seen that the PDF of Z overlaps with that of X (or Y) sometimes, and deviates from each other for some other times. We now investigate the observation in the following subsections.

A. Probability of deviation occurrence

Denote $\mu_{\min} = \min\{\mu_X, \mu_Y\}$ and set intervals

$$P\{\mu_X - t_{X,\frac{\alpha}{2}} \le X \le \mu_X + t_{X,\frac{\alpha}{2}}\} = 1 - \alpha \quad (5)$$

$$P\{\mu_Y - t_{Y,\frac{\beta}{2}} \le Y \le \mu_Y + t_{Y,\frac{\beta}{2}}\} = 1 - \beta. \quad (6)$$

where $t_{X,\frac{\alpha}{2}}$ and $t_{Y,\frac{\beta}{2}}$ are quantiles for the distributions of X and Y. If the following inequality holds, the deviation will occur with the confidence level of $(1-\frac{\alpha}{2})(1-\frac{\beta}{2})$.

$$\delta \le t_{X,\frac{\alpha}{2}} + t_{Y,\frac{\beta}{2}} \quad (7)$$

where $\delta = |\mu_X - \mu_Y|$.

Since a multiple of standard deviation in practice is usually chosen at the intervals (e.g., the three-sigma rule), we give the inequality in terms of a multiple of standard deviation

$$\begin{aligned} \delta &\le m(\mu_{\min} + \delta)^{\frac{1}{2}} + m(\mu_{\min})^{\frac{1}{2}} \\ &= m^2 + 2m(\mu_{\min})^{\frac{1}{2}} \end{aligned} \quad (8)$$

where m is the multiple of standard deviation. If $m = 2$ or $m = 3$, the probability of deviation occurrence is respectively 90.25% or 99.40%.

B. Maximum deviation

The maximum of the deviation is reached when $\mu_X = \mu_Y$, i.e., X and Y follow the same normal distribution. To prove the statement, we focus on the discrete case for simplicity, and let two discrete probability distributions be

$$\begin{cases} \mathrm{P}(X = x_1) = p_1,\ \mathrm{P}(X = x_2) = p_2,\ ...,\ \mathrm{P}(X = x_n) = p_n \\ \mathrm{P}(Y = y_1) = q_1,\ \mathrm{P}(Y = y_2) = q_2,\ ...,\ \mathrm{P}(Y = y_n) = q_n \end{cases} \quad (9)$$

and let

$$x_1 \le y_1 \le x_2 \le y_2, ..., \le x_n \le y_n. \quad (10)$$

Suppose $\mu_X \le \mu_Y$. Then

$$\mu_{\min} = \mu_X, \quad (11)$$

and

$$\begin{aligned} &\mu_{\min} = x_1 p_1 \\ &+x_2 p_2(q_2 + q_3 +, ..., +q_n) +, ..., +x_n p_n(q_n) \\ &+y_1 q_1(p_2 + p_3 +, ..., +p_n) +, ..., +y_{n-1} q_{n-1}(p_n) \\ &= x_1 p_1 + x_2 p_2 +, ..., +x_n p_n \\ &-x_2 p_2(q_1) -, ..., -x_n p_n(q_1 + q_2 +, ..., +q_{n-1}) \\ &+y_1 q_1(p_2 + p_3 +, ..., +p_n) +, ..., +y_{n-1} q_{n-1}(p_n) \\ &= \mu_X + \varepsilon \end{aligned} \quad (12)$$

$$\begin{aligned} \varepsilon \;=\; & q_1[p_2(y_1 - x_2) + p_3(y_1 - x_3) +, ..., +p_n(y_1 - x_n)] \\ +\; & q_2[\qquad\qquad p_3(y_2 - x_3) +, ..., +p_n(y_2 - x_n)] \\ & ... \\ +\; & q_{n-1}[\qquad\qquad\qquad\qquad p_n(y_{n-1} - x_n)] \end{aligned} \quad (13)$$

Since $y_j - x_i < 0$ and $p_i \ge 0, q_j \ge 0$ $(1 \le j < i \le n)$, then $\varepsilon < 0$. When $x_1 = y_1, x_2 = y_2, ..., x_n = y_n$, the differences between x_i and y_j will be largest, so ε reaches the minimum value, and the deviation is hence the largest.

Substituting x^2 and y^2 for x and y, the statement about the variance can be similarly proved. Since any two probability distributions can be approximated with suitable values for p's and q's in (5) this property should be true in general, even in the continuous case. Notice from the proof that the conclusion is still valid to two different normal distributions.

C. The approximation of the maximum deviation

Denote $\mu = \mu_X = \mu_Y$, $F(x) = F_X(x) = F_Y(y)$ and $f(x) = f_X(x) = f_Y(y)$. Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the CDF and PDF of the standard normal distribution. The mean of Z at the time that $\mu_X = \mu_Y$ is

$$\begin{aligned}\mu_Z^* &= \int_{-\infty}^{\infty} z f_Z(z)dz \\ &= \int_{-\infty}^{\infty} z[2f(z) - 2f(z)F(z)]dz \\ &= 2\mu - 2\int_{-\infty}^{\infty} f(z)F(z)dz \\ &= 2\mu - 2\int_{-\infty}^{\infty} (\mu^{\frac{1}{2}}z + \mu)\mu^{-\frac{1}{2}}\phi(z)\Phi(z)d\mu^{\frac{1}{2}}z \\ &= 2\mu - 2\mu^{\frac{1}{2}}\int_{-\infty}^{\infty} z\phi(z)\Phi(z)dz - 2\mu\int_{-\infty}^{\infty}\phi(z)\Phi(z)dz \\ &= \mu - 2a\mu^{\frac{1}{2}}\end{aligned} \tag{14}$$

where $a = \int_{-\infty}^{\infty} z\phi(z)\Phi(z)dz \approx 0.2821$ is constant (the value is numerically approximated using the Simpson's rule). The maximum deviation for the means thus is

$$\Delta_\mu^* = \mu_Z - \mu = 2a\mu^{\frac{1}{2}} \tag{15}$$

Similarly, when $\mu_X = \mu_Y$ the variance of Z and the maximum deviation for the variances are

$$\begin{aligned}(\sigma_Z^2)^* &= \int_{-\infty}^{\infty} z^2 f_Z(z)dz - \left[\int_{-\infty}^{\infty} z f_Z(z)dz\right]^2 \\ &= \mu - 4a^2\mu\end{aligned} \tag{16}$$

$$\Delta_{\sigma^2}^* = 4a^2\mu \tag{17}$$

IV. An approximation method of the mean and variance of Z

In the exact solutions of the mean and variance of Z (i.e., Equation 2 and 3), there are several integrals that impose excessive computational burden and make the method inefficient for applications, in particular, real-time ones. The following approximation method is proposed to overcome this difficulty.

To visualize changing trends of μ_Z with different δ, we first fix $\mu_{\min} = \mu_{\min}(x)$ at a certain value and increase $\mu_{\max} = \max\{\mu_X, \mu_Y\} = \mu_{\max}(x)$ with slope $\mu_{\max}(x)/x = 1$, and then generate curves of $\mu_Z(x)$ corresponding to different $\delta(x)$ using Equation 2 (in which Simpson's rule is utilized to calculate the integrals); see Figure 2 for the trends. An exponentially decay tail is observed, and the following exponential decay equation can be used to fix the shape:

$$f(x, \mu_{\min}(x)) = \mu_{\min}(x) + g_1 \exp\{g_2 x\} \tag{18}$$

where g_1 and g_2 are coefficients.

We assign different values of $\mu_{\min}$ to get more values of g_1 and g_2. Here we take the value from 10 to 10000 and see Table I for the results. Furthermore, the below power forms are employed to fit the relation between $\mu_{\min}$ and g_1, $\mu_{\min}$ and g_2 (see Table II for the fitting results).

$$g_1 = b_1(\mu_{\min})^{\frac{1}{2}} \tag{19}$$

$$g_2 = b_2(\mu_{\min})^{b_3} \tag{20}$$

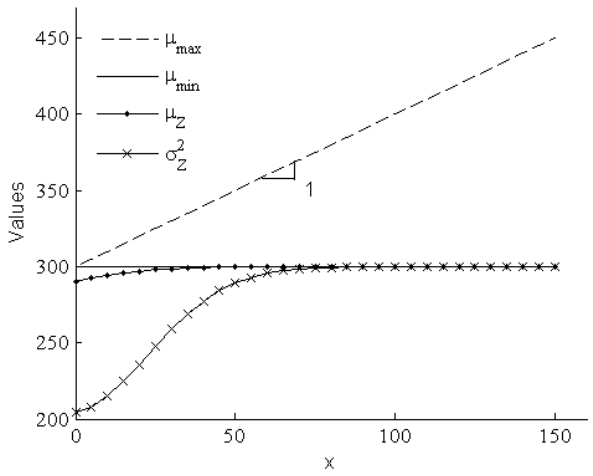


Figure 2. Shapes of μ_Z and σ_Z^2

where $b_1 = -0.5959$, $b_2 = -0.8402$ and $b_3 = -0.4519$. Note that the resulting coefficients contribute to fitting but do not have physical meanings.

Since the slope of $\mu_{\max}(x)$ is equivalent to 1, we have $\delta(x) = x$. Consequently, the approximation of μ_Z is

$$\mu_Z \approx \mu_{\min} + b_1(\mu_{\min})^{\frac{1}{2}} \exp\{b_2(\mu_{\min})^{b_3}\delta\} \tag{21}$$

Likewise, we make use of Sigmoid function to approximate the "S" shape variance of Z (The shape can be observed in Figure 2) as

$$\begin{aligned}\sigma_Z^2 \approx \Delta_{\sigma^2}^* &\left\{1 - \exp\left[-\left(\frac{\mu_Z - \mu_{\min}}{h_1}\right)^{h_2}\right]\right\} + (\sigma_Z^2)^* \\ &\begin{cases} h_1 = c_1(\mu_{\min})^{c_2} \\ h_2 = c_3(\mu_{\min})^{c_4} + c_5, \end{cases}\end{aligned} \tag{22}$$

where $c_1 = 1.911$, $c_2 = 0.4986$, $c_3 = -1.672$, $c_4 - 0.4177$ and $c_5 = 1.996$. Fitting of h_1 and h_2 and goodness of fit of these coefficients are presented Table III and II.

To test the accuracy of the approximations, we calculate the absolute error and the relative error in the range of $\mu_{\min} \in [100, 10000]$ and $\delta \le m^2 + 2m(\mu_{\min})^{\frac{1}{2}}, m = 4$ (See Figure 3). The maximum absolute and relative error of the approximation of μ_Z are less than 4.5 and 3.5×10^{-3}, respectively. The approximation errors of σ_Z^2 are less than 2 and 1.2×10^{-3}. Although there are increasing trends with the growth of means in the absolute errors, the trends do not appear in the relative errors. Therefore, we can say that the approximations of μ_Z and σ_Z^2 give high accuracies in the range from 10 to 10000. Despite the limited range of the approximation, more importance of the work is the method provided. One can modify the values assigned to $\mu_{\min}$ as

Table I
FITTING TO g_1 AND g_2 AT DIFFERENT POINTS OF $\mu_{\min}$

$\mu_{\min}$	g_1	95% confidence bounds	g_2	95% confidence bounds
10	-1.818	$(-1.829, -1.808)$	-0.2949	$(-0.2976, -0.2921)$
50	-4.134	$(-4.164, -4.105)$	-0.1476	$(-0.1492, -0.1460)$
100	-5.879	$(-5.919, -5.838)$	-0.1073	$(-0.1084, -0.1062)$
200	-8.349	$(-8.404, -8.295)$	-0.0774	$(-0.07814, -0.07666)$
500	-13.26	$(-13.33, -13.18)$	-0.04983	$(-0.05024, -0.04942)$
900	-17.82	$(-17.91, -17.73)$	-0.03743	$(-0.03771, -0.03716)$
1400	-22.25	$(-22.36, -22.15)$	-0.03015	$(-0.03036, -0.02995)$
2000	-26.62	$(-26.73, -26.50)$	-0.02530	$(-0.02546, -0.02514)$
3000	-32.62	$(-32.75, -32.49)$	-0.02072	$(-0.02084, -0.02060)$
5000	-42.15	$(-42.30, -41.99)$	-0.01609	$(-0.01618, -0.01601)$
7000	-49.89	$(-50.06, -49.72)$	-0.01362	$(-0.01369, -0.01356)$
10000	-59.65	$(-59.84, -59.46)$	-0.01141	$(-0.01146, -0.01136)$

Table II
GOODNESS OF FIT OF COEFFICIENTS

Equation	Coefficient	Value	95% confidence bounds	SSE*	RMSE**
$g_1 = b_1(\mu_{\min})^{\frac{1}{2}}$	b_1	-0.5959	$(-0.5966, -0.5951)$	3.901×10^{-2}	5.955×10^{-2}
$g_2 = b_2(\mu_{\min})^{b_3}$	b_2	-0.8402	$(-0.8691, -0.8113)$	4.878×10^{-5}	2.209×10^{-3}
	b_3	-0.4519	$(-0.4620, -0.4417)$		
$h_1 = c_1(\mu_{\min})^{c_2}$	c_1	1.911	$(1.904, 1.918)$	2.732×10^{-2}	5.226×10^{-2}
	c_2	-0.7414	$(-0.7765, -0.7063)$		
$h_2 = c_3(\mu_{\min})^{c_4} + c_5$	c_3	-1.672	$(-1.72, -1.625)$	7.403×10^{-5}	2868×10^{-3}
	c_4	-0.4177	$(-0.4307, -0.4047)$		
	c_5	1.996	$(1.99, 2.002)$		

* Sum of Squares Due to Error
** Root Mean Squared Error

Table III
FITTING TO h_1 AND h_2 AT DIFFERENT POINTS OF $\mu_{\min}$

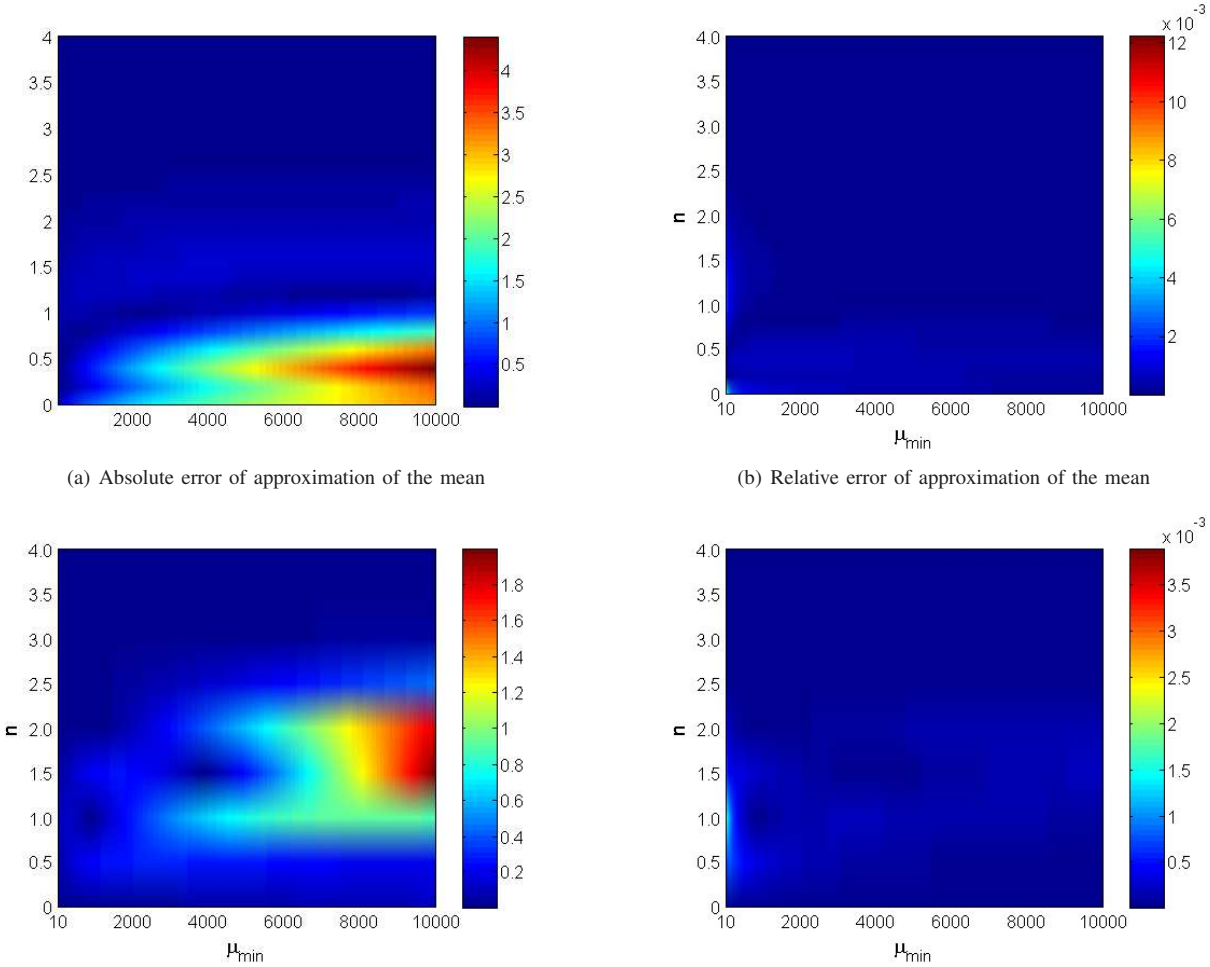
$\mu_{\min}$	h_1	95% confidence bounds	h_2	95% confidence bounds
10	5.901	$(5.866, 5.936)$	1.358	$(1.342, 1.374)$
50	13.48	$(13.45, 13.51)$	1.664	$(1.654, 1.674)$
100	19.04	$(19.01, 19.07)$	1.75	$(1.744, 1.757)$
200	26.88	$(26.85, 26.9)$	1.815	$(1.81, 1.819)$
500	42.38	$(42.36, 42.4)$	1.874	$(1.872, 1.876)$
900	56.78	$(56.77, 56.8)$	1.901	$(1.899, 1.902)$
1400	70.76	$(70.75, 70.78)$	1.916	$(1.915, 1.917)$
2000	84.53	$(84.51, 84.54)$	1.927	$(1.926, 1.928)$
3000	103.5	$(103.5, 103.5)$	1.937	$(1.936, 1.937)$
5000	133.5	$(133.5, 133.5)$	1.947	$(1.946, 1.947)$
7000	157.9	$(157.9, 157.9)$	1.952	$(1.952, 1.952)$
10000	188.7	$(188.7, 188.7)$	1.957	$(1.956, 1.957)$

needed to obtain an approximation for a larger or another range.

V. CONCLUSIONS

This paper approximated the minimum distribution of two normally distributed variables each with the same mean and variance. Using the approximation, the computational efficiency is improved greatly and it is critical to applications in engineering. Since the normal distribution with the same mean and variance is good approximation of Poisson distribution, the approximation can be extended in calculating the minimum of two Poisson distributed variables.

In the method, one variable was fixed first, and then the values of the minimum corresponding to different values of the other variable were fitted; by changing values of the first variable, groups of parameters were obtained and fitted; an approximation was derived consequently. The test results showed that the approximation gave high accuracy in the range from 10 to 10000. More importance is the method; one can get approximations for other ranges by modifying the parameters. In addition, the occurrence probability of deviation between the two normal distributions and the minimum distribution was presented; It was also proved that the the maximum deviation was reached when the two



(a) Absolute error of approximation of the mean

(b) Relative error of approximation of the mean

(c) Absolute error of approximation of the variance

(d) Relative error of approximation of the variance

Figure 3. Approximation errors of the mean and variance of Z

normal distributions intersected. There feathers benefit the understanding of the minimum distribution.

ACKNOWLEDGMENT

This research has been funded by the Fundamental Research Funds for the Central Universities (2012JBM064) and the National Natural Science Foundation of China (71131001-2). The authors are also grateful to Wang L. for checking the phrasing patiently.

REFERENCES

[1] Gumbel E J. Statistical theory of extreme values and some practical applications. *Applied mathematics series*, 1954.

[2] Alfredo A, Wilson T. Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering (2 edition). *Wiley*, 2006.

[3] Aroian L A, Taneja V S, Cornwell L W. Mathematical Forms of the Distribution of the Product of Two Normal Variables. *Communications in Statistics - Theory and Methods*, 7: 164-172, 1978.

[4] Cornwell W, Aroian L A, Taneja V S. Numerical Evaluation of the Distribution of the Product of Two Normal Variables. *Journal of Statistical Computation and Simulation*, 7: 123-131 ,1978.

[5] Glen A G, Leemis L M, Drew J H. Computing the distribution of the product of two continuous random variables. *Computational Statistics and Data Analysis*,44: 451-464,2004.