On an Order Relation Between Distributions with Applications to Interval Estimation

R. Alt[1], I. Dakov[2], and S. Markov[2]

[1]Laboratoire LIP6, Universite Pierre et Marie Curie
4 Place Jussieu, 75005 Paris, France
[2]Section "Biomathematics"
Inst. of Mathematics and Computer Sci.
Bulgarian Academy of Sciences
"Acad. G. Bonchev" st., block 8,
BG-1113 Sofia, Bulgaria

1 Introduction

A problem which appears quite often in the validation of the results of numerical algorithms using intervals is the branching based on the comparison of two intervals when these intervals have a noon-empty intersection. This paper deal with some possibility of ordering intervals with some probability and a technique for the computation of this probability.

Let f be a density function (i.e., a Lebesgue integrable function on $\mathbb{R}$ such that $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$) and let $F(x) = \int_{-\infty}^{x} f(t)dt$ be the distribution (function) of f.

Given two independant real random variables ξ, η with known densities (or distributions) we want to compute the probability $P(\xi > \eta)$. This may be useful whenever the random variables represent numbers containing stochastic errors as is the case with stochastic numbers [1–4].

Denote by f_ξ, g_η the density functions of the random variables ξ, η resp. If both densities f_ξ, g_η have intervals as support sets $A, B \subset \mathbb{R}$, that is:

i) $f_\xi \neq 0$ for $\xi \in A$, and $f_\xi = 0$ for $\xi \notin A$;

ii) $g_\eta \neq 0$ for $\eta \in B$, and $g_\eta = 0$ for $\eta \notin B$,

and if the support sets do not intersect ($A \cap B \neq \emptyset$), then we have that $P(\xi > \eta)$ is either 0 or 1, depending on whether $A > B$ or $A < B$. In these cases one can speak of an order relation between distributions, that is a relation of the form $f_\xi < g_\eta$, resp. $f_\xi > g_\eta$. In the general case, when $0 < P(\xi > \eta) < 1$ the number $P(\xi > \eta)$ can serve as a measure (indicator) for such an ordering. For this reason we shall denote $M(f_\xi, g_\eta) = P(\xi > \eta)$.

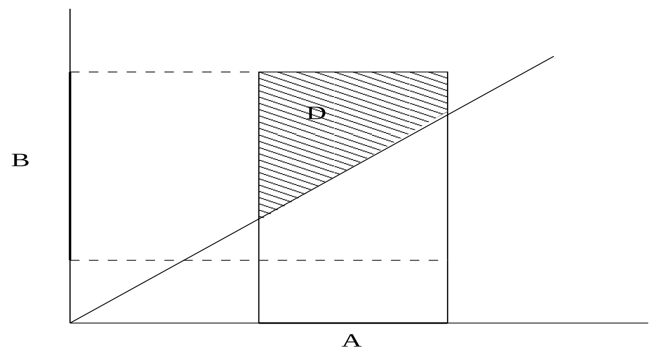

Figure 1: *domain definition for Pr(ζ < 0)*

In what follows we investigate some rules for the computation of $M(f_\xi, g_\eta)$ especially suitable for the case when the support sets of the densities are (compact) intervals. A possible way to compute $M(f_\xi, g_\eta)$ is classically the following. The definition domain of $\Pr(\zeta < 0)$ is represented in figure (1), so we have:

$$\begin{aligned} M(f_\xi, g_\eta) &= P(\xi > \eta) = P(\zeta > 0) = 1 - P(\zeta \leq 0) \\ &= 1 - \int_{-\infty}^{0} h_\zeta(\zeta) d\zeta = 1 - \int_{z=-\infty}^{0} \int_{t=-\infty}^{\infty} f_\xi(z+t) g_\eta(t) dz \; dt, \quad (1) \end{aligned}$$

where $f_\xi(x)$ is the density of ξ, and $g_\eta(x)$ is the density of η.
Thus formula (1) can be used for the computation of $M(f_\xi, g_\eta) = P(\xi > \eta)$. However, we show here that an easier technique can be used in the case when the support sets of the densities are intervals.

So, in the present work a simple method is proposed for the computation of the probability $P(\xi > \eta)$, where ξ, η are two real random variables with known densities. The method is illustrated for some familiar densities Gaussian, uniform).

2 Technique

Let $\mathbb{R}^* = \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$. Let $B = [\underline{b}, \overline{b}]$, $\underline{b} \leq \overline{b}$, $\underline{b}, \overline{b} \in \mathbb{R}^*$. The probability $P(\xi \in B)$ for a random variable with a density f to belong to the interval B is given by $P(\xi \in B) = F(\overline{b}) - F(\underline{b}) = \int_{\underline{b}}^{\overline{b}} f_\xi(t) dt$; note that $F(-\infty) = 0$ and $F(\infty) = 1$.

Proposition. *For any two random variables ξ, η with given density functions f_ξ, g_η and any integer $n \geq 0$ and system of real numbers $t_1, \ldots, t_n \in \mathbb{R}$ and $t_i < t_{i+1}$ (for $n = 0$ the system is considered empty), we have:*

$$M(f_\xi, g_\eta) = \sum_{k \geq l, k, l = 1}^{n+1} p_{kl} \int_{t_{k-1}}^{t_k} f_\xi(t) dt \int_{t_{l-1}}^{t_l} g_\eta(t) dt, \quad (2)$$

wherein

$$p_{kl} = \begin{cases} 0, & k < l, \\ 1, & k > l. \end{cases} \tag{3}$$

In the case $k = l$ the value of p_{kk} depends on the details of f_ξ and g_η.

Proof. Denote $A_1 = [-\infty, t_1], A_i = [t_{i-1}, t_i], i = 1, \ldots, n, A_{n+1} = [t_n, \infty]$. Applying the conditional probabilities formula for the division of $\mathbb{R}$ we can write:

$$\begin{aligned} M(f_\xi, g_\eta) &= P(\xi > \eta) = \sum_{k,l=1}^{n+1} P(\xi > \eta, \xi \in A_k, \eta \in A_l) \\ &= \sum_{k,l=1}^{n+1} P(\xi > \eta \mid \xi \in A_k, \eta \in A_l) P(\xi \in A_k) P(\eta \in A_l) \\ &= \sum_{k,l=1}^{n+1} p_{kl} P(\xi \in A_k) P(\eta \in A_l), \end{aligned}$$

where $p_{kl} = P(\xi > \eta \mid \xi \in A_k, \eta \in A_l) = \{0, \text{ if } k < l; 1, \text{ if } k > l\}$.

In the case $k = l$, that let us call $\tilde{f}_\xi$ and $\tilde{g}_\eta$ the densities of ξ and η on A_k. (Note that they are not the restrictions of f_ξ and g_η to A_k). Then we have:

$$p_{kk} = \int_{x=t_{k-1}}^{t_k} \int_{y=t_{k-1}}^{x} \tilde{f}_\xi(x) \tilde{g}_\eta(y) dx dy \tag{4}$$

This implies the proposition.

The above formula can be considered as a discretisation of (1) but it has the advantages that the integrals (4) are defined on the same sub-interval for the two variables and are thus easier to compute. Moreover the considered sub-intervals can be small and it may be easy to obtain bounds for the coefficients p_{kk}.

Three obvious cases when the value of M is easily determined can be retreived with this formula:

For any two random variables ξ, η with given density functions f, g, resp., we have

a) If $f \equiv g$, then $M(f, g) = 1/2$;

b) In the case when $f(x) = 0$ for every $x \le t$, and $g(x) = 0$ for every $x \ge s$, $t \ge s$, then $M(f, g) = 1$;

c) In the case when $g(x) = 0$ for every $x \le t$, and $f(x) = 0$ for every $x \ge s$, $t \ge s$, then $M(f, g) = 0$.

These cases are retreived under appropriate division $\{t_i\}$ on $\mathbb{R}$. Namely, in the case a) take $\mathbb{R}$ as a single interval (empty division set, $n = 0$); in the cases b), c) take $n = 1$, $t_1 = (s + t)/2$.

In what follows the preceding technique based on formula (2) is applied to the computation of the measure M in the cases of some well-known distributions.

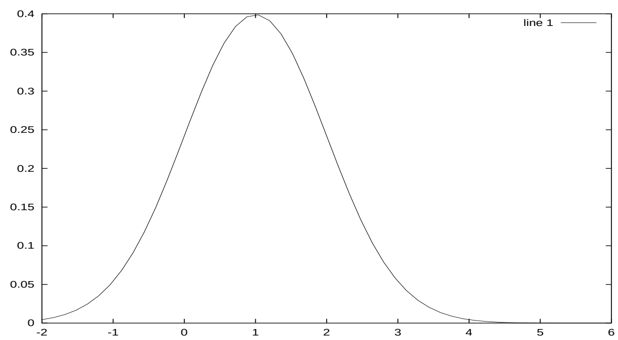

Figure 2: *two Gaussian distributions*

3 Uniform Distributions

Let us consider two random variables ξ and η defined on two intersecting intervals A and B. $A = [\underline{a}, \overline{a}]$ and $B = [\underline{b}, \overline{b}]$ with $\overline{a} > \underline{b}$. Take $n = 3$ and $t_1 = \underline{b}$ and $t_2 = \overline{a}$. With the notations of (3) we have obviously: $p_{12} = p_{13} = p_{23} = 0$ and $p_{21} = p_{31} = p_{32} = 1$. Concerning p_{11}, p_{22}, p_{33} we are in the case of both ξ and η belonging to the same sub-interval with a same constant probability density. So $p_{11} = p_{22} = p_{33} = 1/2$. Note that in fact only p_{22} has an interest as in formula (2) the probabilities $P(\eta \in A_1)$ and $P(\xi \in A_3)$ are null.

4 Gaussian Distributions

Let us consider the same intersecting intervals as above but with two different Gaussian distributions f_ξ, g_η on each interval, having mean values m_1, m_2, resp., and variances σ_1^2, σ_2^2, resp. Thus the density functions of ξ and η are: $f_\xi(x) = (2\pi\sigma_1^2)^{-1/2} e^{-(x-m_1)^2/(2\sigma_1^2)}$, $g_\eta(x) = (2\pi\sigma_2^2)^{-1/2} e^{-(x-m_2)^2/(2\sigma_2^2)}$. Assume $m_1 \leq m_2$. Take $n = 2$, t_1 and t_2 is the abscissas of the points common to $f_\xi(x)$ and $g_\eta(x)$, and denote $A_1 = [-\infty, t_1]$, $A_2 = [t_1, t_2]$, and $A_3 = [t_2, \infty]$. As before formula (3) gives $p_{kl} = 0$ for$k < l$ and $p_{kl} = 1$ for $k > l$. Concerning the coefficients $p_{kk}, k = 1, 2, 3$ they depend on the values of $m_1, m_2, \sigma_1, \sigma_2$ and they have to be computed with the integral of formula (4). As example let us here compute p_{11}. By hypothese ξ and η are in A_{11} then $Pr(\xi) \in A_1 = 1$ and $Pr(\eta) \in A_1 = 1$. So let us call: $S_1 = \int_{-\infty}^{t_1} f_\xi(t)dt$ and $S_2 = \int_{-\infty}^{t_1} g_\eta(t)dt$. Then $\tilde{f}_\xi = f_\xi/S_1$ and $\tilde{g}_\eta = g_\eta/S_2$ and $p_{11} = (\int_{x=-\infty}^{t_1} \int_{y=-\infty}^{x} f_\xi(x) g_\eta(y) dxdy)/(S_1 S_2)$.

Special case 1. $m_1 = m_2$. Then $(t_1 + t_2)/2 = m_1$. The problem is symmetric and $p_{11} = 1 - p_{33}$.

Special case 2. $\sigma_1 = \sigma_2 = \sigma$. In this case $t_1 = t_2$ and $A_{22} = \varnothing$.

Similar formulae have been deduced for the Beta distribution.

5 A Note on Applications

These results can be be used for treating branchings using comparisons of intervals when these intervals have a non empty intersection. They can also be applied to the problems of interpolation and approximation in the case of interval (uncertain but bounded) data with given densities (distributions). Thus this approach may find useful applications in mathematical modelling situations. On the other hand this approach contributes to the arithmetic theory of stochastic numbers [3].

References

[1] R. Alt and S. Markov, "On the Algebraic Properties of Stochastic Arithmetic. Comparison to Interval Arithmetic", In: W. Krämer and J. Wolff von Gudenberg (eds)., *Scientific Computing, Validated Numerics, Interval Methods*, Kluwer, 2001, pp. 331–341.

[2] R. Alt and J. Vignes, "Validation of Results of Collocation Methods for ODEs with the CADNA Library", *Appl. Numer. Math.*, 1996, Vol. 20, pp. 1–21.

[3] J. M. Chesneaux and J. Vignes, "Les fondements de l'arithmétique stochastique", *C.R Acad. Sci.*, Paris, Sér. I, Math., 1992, Vol. 315, pp. 1435–1440.

[4] J. Vignes and R. Alt, "An Efficient Stochastic Method for Round-Off Error Analysis", In: *Accurate Scientific Computations*, L.N.C.S. 235, Springer, 1985, pp. 183–205.