EXACT KOLMOGOROV AND TOTAL VARIATION DISTANCES BETWEEN SOME FAMILIAR DISCRETE DISTRIBUTIONS

JOSÉ A. ADELL AND P. JODRÁ

We give exact closed-form expressions for the Kolmogorov and the total variation distances between Poisson, binomial, and negative binomial distributions with different parameters. In the Poisson case, such expressions are related with the Lambert W function.

1. Introduction

Estimates of the closeness between probability distributions measured in terms of certain distances, particularly, the Kolmogorov and the total variation distances are very common in theoretical and applied probability. Usually, the results refer to upper estimates of those distances, even sharp upper bounds in some sense. As far as we know, only a few exceptions deal with exact formulae (see, e.g., Kennedy and Quine [5], where the exact total variation distance between binomial and Poisson distributions is given for small values of the success parameter of the binomial). Although numerical computations seem to be unavoidable, exact expressions are only useful if they are easy to handle.

The aim of this note is to provide exact closed-form expressions for the Kolmogorov and the total variation distances between Poisson, binomial, and negative binomial distributions with different parameters. In many occasions, these distances appear as ingredients to estimate other distances in more complex situations (see, e.g., Ruzankin [8]). On the other hand, it is interesting to observe that, in the Poisson case, such exact formulae involve the Lambert W function. This function, for which efficient numerical procedures of evaluation are known, has many applications in pure and applied mathematics (for more details, see Corless et al. [3], Barry et al. [2], and the references therein).

Denote by $\mathbb{N}$ the set of nonnegative integers and by $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$. Given two $\mathbb{N}$-valued random variables X and Y, the Kolmogorov and the total variation distances between them are respectively defined by $d_K(X, Y) := \sup_{k \in \mathbb{N}} |P(X \geq k) - P(Y \geq k)|$ and

$d_{\mathrm{TV}}(X,Y) := \sup_{A \subseteq \mathbb{N}} |P(X \in A) - P(Y \in A)|$. We denote by

$$f(k) := \frac{P(X = k)}{P(Y = k)}, \quad k \in \mathbb{N}, \qquad \left(\frac{a}{0} := \infty,\ a \geq 0\right). \tag{1.1}$$

All of the examples in the following section rely upon the following easy result.

THEOREM 1.1. *If the function* $f(\cdot)$ *is nondecreasing, then*

$$d_K(X,Y) = d_{\mathrm{TV}}(X,Y) = P(X \geq \ell) - P(Y \geq \ell), \tag{1.2}$$

where $\ell := \inf\{k \in \mathbb{N} : f(k) \geq 1\}$.

Proof. Since $f(\cdot)$ is nondecreasing, we have that $\{k \in \mathbb{N} : P(X = k) \geq P(Y = k)\} = \{\ell, \ell+1,\ldots\}$. This readily implies the statements in Theorem 1.1. □

2. Examples

Poisson, binomial, and negative binomial distributions are among the most widely used discrete distributions in modelling different phenomena. In this section, we give exact distances for these distributions and recall some related upper estimates available in the literature.

2.1. Poisson distributions.   For any $t > 0$, let $N(t)$ be a random variable having the Poisson distribution with mean t, that is,

$$P(N(t) = k) := e^{-t}\frac{t^k}{k!}, \quad k \in \mathbb{N}. \tag{2.1}$$

Some upper bounds for the total variation distance between two Poisson distributions with different means are the following:

$$d_{\mathrm{TV}}(N(t+x),N(t)) \leq \min\left\{1 - e^{-x}, \int_t^{t+x} P(N(u) = \lfloor u \rfloor)\,du\right\}$$

$$\leq \int_t^{t+x} P(N(u) = \lfloor u \rfloor)\,du \leq \min\left\{x, \sqrt{\frac{2}{e}}(\sqrt{t+x} - \sqrt{t})\right\}, \quad t,x \geq 0, \tag{2.2}$$

where $\lfloor x \rfloor$ stands for the integer part of x. The first upper bound in (2.2) is given in Adell and Lekuona [1, Corollary 3.1], the second in Ruzankin [8, Lemma 1], while the third can be found in Roos [7, formula (5)]. On the other hand, the Poisson-gamma relation states (cf. Johnson et al. [4, page 164]) that

$$P(N(t) \leq n) = \int_t^{\infty} P(N(u) = n)\,du, \quad n \in \mathbb{N},\ t \geq 0. \tag{2.3}$$

For any $x \geq 0$, we denote by $\lceil x \rceil$ the ceiling of x, that is, $\lceil x \rceil := \inf\{k \in \mathbb{N} : k \geq x\}$. Concerning Poisson distributions, we enunciate the following.
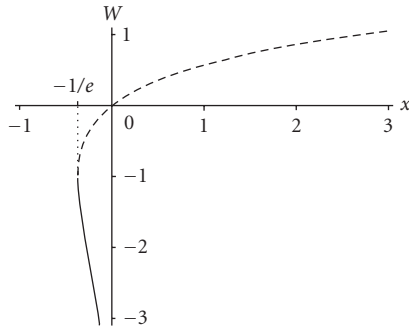
Figure 2.1. The two real branches of $W(x)$. Dashed line: $W_{-1}(x)$; dotted line: $W_0(x)$.

PROPOSITION 2.1. *For any $t > 0$ and $x > 0$, we have*

$$d_K\left(N(t+x),N(t)\right) = d_{\mathrm{TV}}\left(N(t+x),N(t)\right) = \int_t^{t+x} P\left(N(u) = \ell - 1\right) du, \qquad (2.4)$$

where

$$\lceil t \rceil \le \ell := \ell(t,x) = \left\lceil \frac{x}{\log(1+x/t)} \right\rceil \le \lceil t + x \rceil. \qquad (2.5)$$

Proof. Fix $t > 0$ and $x > 0$. Observe that the function

$$f(k) := \frac{P(N(t+x) = k)}{P(N(t) = k)} = e^{-x}\left(1+\frac{x}{t}\right)^k, \quad k \in \mathbb{N} \qquad (2.6)$$

is increasing and that $\inf\{k \in \mathbb{N} : f(k) \ge 1\} = \ell$, as defined in (2.5). Therefore, (2.4) follows from Theorem 1.1 and (2.3). The first inequality in (2.5) follows from the well-known inequality $\log(1 + y) \le y$, $y \ge 0$, while the second follows from the fact that $(t + y)\log(1 + (y/t)) \ge y$, $y \ge 0$. The proof is complete. $\qquad \square$

In view of Proposition 2.1, it may be of interest to characterize the sets

$$A_\ell := \{(t,x) : t > 0, \ x > 0, \ \ell(t,x) = \ell\}, \quad \ell \in \mathbb{N}^*. \qquad (2.7)$$

To this end, we consider the Lambert W function (see Figure 2.1), defined as the solution to the equation

$$W(x)e^{W(x)} = x, \quad x \ge -\frac{1}{e}. \qquad (2.8)$$

For $-1/e \le x < 0$, there are two possible real branches of $W(x)$. We will only be interested in the branch taking on values in $(-\infty, -1]$, denoted in the literature by $W_{-1}(x)$. It is known that $W_{-1}(-1/e) = -1$, $W_{-1}(x)$ is decreasing and that $W_{-1}(x) \to -\infty$ as $x \to 0$. A review of the history, theory and applications of the Lambert W function may be found in Corless et al. [3] and Barry et al. [2].
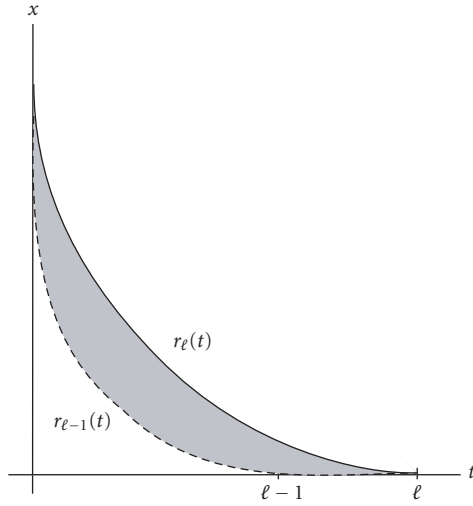
Figure 2.2. Picture of A_ℓ as the shadowed region.

Let $k \in \mathbb{N}$ and $t > 0$. We consider the function

$$g_{k,t}(x) := e^{-x}\left(1 + \frac{x}{t}\right)^k, \quad x \geq 0. \tag{2.9}$$

The following properties are easy to check. The equation $g_{k,t}(x) = 1$ has $x = 0$ as the unique solution if $k \leq t$, and has one positive solution, together with the null solution, if $k > t$. Denote by $r_k(t)$ the largest solution to the equation $g_{k,t}(x) = 1$. Since $g_{k+1,t}(x) > g_{k,t}(x), x > 0, k \in \mathbb{N}$, we see that

$$r_0(t) = \cdots = r_{\lfloor t \rfloor}(t) = 0 < r_{\lfloor t \rfloor + 1}(t) < r_{\lfloor t \rfloor + 2}(t) < \cdots. \tag{2.10}$$

On the other hand, by (2.8), (2.9), and the aforementioned properties of $W_{-1}(x)$, it can be verified that for any $k \in \mathbb{N}^*$ we have

$$r_k(t) = \begin{cases} -kW_{-1}\left(-\dfrac{t}{k}\, e^{-t/k}\right) - t, & 0 < t < k \\ 0, & k \leq t. \end{cases} \tag{2.11}$$

A graphical representation of these functions is given in **Figure 2.2** (see also the remark at the end of this note). We state the following.

PROPOSITION 2.2. *Let A_ℓ be as in (2.7), $\ell \in \mathbb{N}^*$. Then,*

$$A_\ell = \{(t,x) : t > 0,\ r_{\ell-1}(t) < x \leq r_\ell(t)\}. \tag{2.12}$$

Proof. Let $t > 0$ and $x > 0$. By (2.5) and (2.9), $\ell(t,x) = \ell \in \mathbb{N}^*$ if and only if $g_{\ell-1,t}(x) < 1 \leq g_{\ell,t}(x)$. By (2.9) and (2.10), this is equivalent to $r_{\ell-1}(t) < x \leq r_\ell(t)$. The proof is complete. $\qquad\square$

2.2. Binomial distributions. Let $n \in \mathbb{N}^*$, $0 < p < 1$, and $q := 1 - p$. Denote by $S_n(p)$ a random variable having the binomial distribution with parameters n and p, that is,

$$P(S_n(p) = k) := \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n. \tag{2.13}$$

The well-known binomial-beta relation (cf. Johnson et al. [4, page 117]) reads as

$$P(S_n(p) \geq k) = n \int_0^p P(S_{n-1}(u) = k - 1) du, \quad n \in \mathbb{N}^*, \; k = 1, \ldots, n. \tag{2.14}$$

Let $0 < p < 1$ and $0 < x < 1 - p$. Roos [6, formula (15)] has given the upper bound

$$d_{\mathrm{TV}}(S_n(p+x), S_n(p)) \leq \frac{\sqrt{e}}{2} \frac{\tau(x)}{\left(1 - \tau(x)\right)^2}, \tag{2.15}$$

where

$$\tau(x) := x \sqrt{\frac{n+2}{2p(1-p)}}, \tag{2.16}$$

provided that $\tau(x) < 1$. Estimate (2.15) is a particular case of much more general results referring to binomial approximation of Poisson binomial distributions obtained by Roos [6]. With respect to binomial distributions, we give the following.

PROPOSITION 2.3. *Let $n \in \mathbb{N}^*$, $0 < p < 1$, and $0 < x < q := 1 - p$. Then,*

$$d_K(S_n(p+x), S_n(p)) = d_{\mathrm{TV}}(S_n(p+x), S_n(p)) = n \int_p^{p+x} P(S_{n-1}(u) = \ell - 1) du, \tag{2.17}$$

where

$$\lceil np \rceil \leq \ell := \ell_p(n,x) = \left\lceil \frac{-n \log(1 - x/q)}{\log(1 + x/p) - \log(1 - x/q)} \right\rceil \leq \lceil n(p+x) \rceil. \tag{2.18}$$

Proof. Since the logarithmic function is concave, we have

$$p \log\left(1 + \frac{x}{p}\right) + q \log\left(1 - \frac{x}{q}\right) \leq \log 1 = 0, \quad 0 \leq x < q. \tag{2.19}$$

This clearly implies the first inequality in (2.18). On the other hand, the function

$$h(x) := (p+x) \log\left(1 + \frac{x}{p}\right) + (q - x) \log\left(1 - \frac{x}{q}\right), \quad 0 \leq x < q \tag{2.20}$$

is nonnegative, because $h(0) = 0$ and $h'(x) \geq 0$, $0 \leq x < q$. The nonnegativity of h implies the second inequality in (2.18). The remaining assertions follow as in proof of Proposition 2.1, replacing the Poisson-gamma relation by (2.14). The proof is complete.
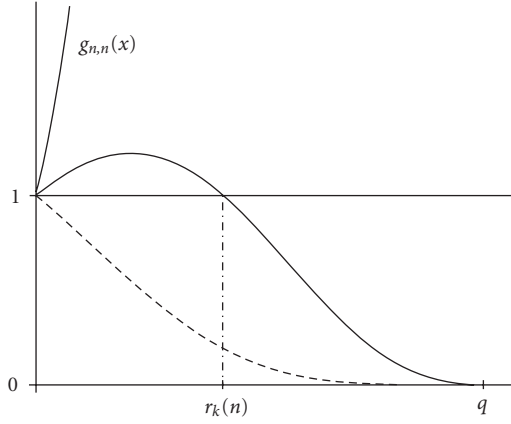
□

Figure 2.3. The functions $g_{k,n}(x)$. Dashed line, if $np < k < n$; dotted line, if $k \le np$.

Let $p \in (0,1)$ be fixed. Recalling the notation in (2.18), we consider the sets

$$B_\ell := \{(n,x) : n \ge \ell, \ 0 < x < q, \ \ell_p(n,x) = \ell\}, \quad \ell \in \mathbb{N}^*. \tag{2.21}$$

For any $k \in \mathbb{N}$ and $n \in \mathbb{N}^*$ with $k \le n$, we define the function (see Figure 2.3)

$$g_{k,n}(x) := \left(1 + \frac{x}{p}\right)^k \left(1 - \frac{x}{q}\right)^{n-k}, \quad 0 \le x < q. \tag{2.22}$$

The equation $g_{k,n}(x) = 1$ has $x = 0$ as the unique solution if $k \le np$ or $k = n$, and has one solution in $(0,q)$, together with the null solution, if $np < k < n$. Denote by $r_k(n)$ the largest solution to the equation $g_{k,n}(x) = 1$ in $[0,q)$. It is easily checked (see Figure 2.3) that

$$r_n(n) = r_0(n) = \cdots = r_{\lfloor np \rfloor}(n) = 0 < r_{\lfloor np \rfloor+1}(n) < \cdots < r_{n-1}(n) < q. \tag{2.23}$$

PROPOSITION 2.4. *Let $p \in (0,1)$ be fixed and let B_ℓ be as in (2.21), $\ell \in \mathbb{N}^*$. Then,*

$$B_\ell = \{\ell\} \times (r_{\ell-1}(\ell),q) \bigcup \left(\ell, \frac{\ell}{p}\right) \times (r_{\ell-1}(n), r_\ell(n)]. \tag{2.24}$$

Proof. Let $n \in \mathbb{N}^*$. For any $n \ge \ell$ and $0 < x < q$, we have from (2.18) that $\ell_p(n,x) = \ell$ if and only if

$$g_{\ell-1,n}(x) < 1 \le g_{\ell,n}(x). \tag{2.25}$$

From (2.22) and (2.23), we have the following. If $n = \ell$, (2.25) is equivalent to $r_{\ell-1}(\ell) < x < q$. If $\ell < n < \ell/p$, (2.25) is equivalent to $r_{\ell-1}(n) < x \le r_\ell(n)$. Finally, if $n \ge \ell/p$, (2.25) has no solution. The proof is complete. $\qquad\square$

2.3. Negative binomial distributions. Let $m \in \mathbb{N}^*$, $0 < p < 1$, and $q := 1 - p$. Let $T_m(p)$ be a random variable such that

$$P(T_m(p) = k) = \binom{m+k-1}{k} p^m q^k, \quad k \in \mathbb{N}. \tag{2.26}$$

The negative binomial-beta relation can be written (cf. Johnson et al. [4, page 210]) as

$$P(T_m(p) \leq k) = (m+k) \int_0^p P(S_{m+k-1}(u) = m-1) du, \quad k \in \mathbb{N}, \tag{2.27}$$

where $S_n(u)$ is defined in (2.13). We will simply state the results referring to negative binomial distributions, because their proofs are very similar to those in the preceding example. The main difference is that relation (2.27) must be used instead of (2.14).

PROPOSITION 2.5. *Let $m \in \mathbb{N}^*$, $p \in (0,1)$, and $0 < x < q := 1 - p$. Then,*

$$d_K(T_m(p), T_m(p+x)) = d_{TV}(T_m(p), T_m(p+x))$$

$$= (m+\ell-1) \int_p^{p+x} P(S_{m+\ell-2}(u) = m-1) du, \tag{2.28}$$

where

$$\left\lceil m \frac{q-x}{p+x} \right\rceil \leq \ell := \ell_p(m,x) = \left\lceil -m \frac{\log(1+x/p)}{\log(1-x/q)} \right\rceil \leq \left\lceil m \frac{q}{p} \right\rceil. \tag{2.29}$$

Let $p \in (0,1)$ be fixed. We denote by

$$C_\ell := \{(m,x) : m \in \mathbb{N}^*, 0 < x < q, \ell_p(m,x) = \ell\}, \quad \ell \in \mathbb{N}^*. \tag{2.30}$$

On the other hand, for any $m \in \mathbb{N}^*$ and $k \in \mathbb{N}$, we consider the function

$$g_{m,k}(x) := \left(1 + \frac{x}{p}\right)^m \left(1 - \frac{x}{q}\right)^k, \quad 0 \leq x < q. \tag{2.31}$$

It turns out that

$$g_{m,0}(x) > g_{m,1}(x) > \cdots > g_{m,k}(x) > \cdots, \quad 0 < x < q, k \in \mathbb{N}. \tag{2.32}$$

The equation $g_{m,k}(x) = 1$ has $x = 0$ as the unique solution if $k = 0$ or if $k \geq mq/p$, and has one solution in $(0,q)$, together with the null solution, if $0 < k < mq/p$. Denote by $r_k(m)$ the largest solution to the equation $g_{m,k}(x) = 1$ in $[0,q)$. By (2.32), we have that

$$\cdots = r_{\lceil mq/p \rceil+1}(m) = r_{\lceil mq/p \rceil}(m) = r_0(m) = 0 < r_{\lceil mq/p \rceil-1}(m) < \cdots < r_1(m) < q. \tag{2.33}$$

With the preceding notations, we state the following.

PROPOSITION 2.6. *Let $p \in (0,1)$ be fixed and let C_ℓ be as in (2.30), $\ell \in \mathbb{N}^*$. Then,*

$$C_\ell = \left(\left\lfloor \frac{p(\ell-1)}{q} \right\rfloor, \left\lfloor \frac{p\ell}{q} \right\rfloor\right] \times (0, r_{\ell-1}(m)) \bigcup \left(\left\lfloor \frac{p\ell}{q} \right\rfloor, \infty\right) \times [r_\ell(m), r_{\ell-1}(m)). \tag{2.34}$$

Final remark 2.7. From a computational point of view, there is a substantial difference in determining the sets A_ℓ, on the one hand, and the sets B_ℓ and C_ℓ, on the other, $\ell \in \mathbb{N}^*$. In the Poisson case, formula (2.11) gives us closed-form expressions for the functions $r_k(t)$, defining the sets A_ℓ, in terms of the Lambert W function. Since this function is implemented in various computer algebra systems—Maple, for instance—the functions $r_k(t)$ can be evaluated in a straightforward manner. In the binomial case, in contrast, we do not know any function, implemented in some computer algebra system, in terms of which the functions $r_k(n)$, defining the sets B_ℓ, could be expressed. In such circumstances, the values $r_k(n)$ must be numerically computed one by one, for each fixed value of the parameters k, n, and p. Similar considerations are valid in the negative binomial case.

Acknowledgment

References

[1] J. A. Adell and A. Lekuona, *Sharp estimates in signed Poisson approximation of Poisson mixtures*, Bernoulli **11** (2005), no. 1, 47–65.

[2] D. A. Barry, J.-Y. Parlange, L. Li, H. Prommer, C. J. Cunningham, and F. Stagnitti, *Analytical approximations for real values of the Lambert W-function*, Mathematics and Computers in Simulation **53** (2000), no. 1-2, 95–103.

[3] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, *On the Lambert W function*, Advances in Computational Mathematics **5** (1996), no. 4, 329–359.

[4] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate Discrete Distributions*, 2nd ed., Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, 1992.

[5] J. E. Kennedy and M. P. Quine, *The total variation distance between the binomial and Poisson distributions*, The Annals of Probability **17** (1989), no. 1, 396–400.

[6] B. Roos, *Binomial approximation to the Poisson binomial distribution: the Krawtchouk expansion*, Theory of Probability and Its Applications **45** (2001), no. 2, 258–272.

[7] ———, *Improvements in the Poisson approximation of mixed Poisson distributions*, Journal of Statistical Planning and Inference **113** (2003), no. 2, 467–483.

[8] P. S. Ruzankin, *On the rate of Poisson process approximation to a Bernoulli process*, Journal of Applied Probability **41** (2004), no. 1, 271–276.

José A. Adell: Departamento de Métodos Estadísticos, Universidad de Zaragoza, 50009 Zaragoza, Spain
E-mail address: adell@unizar.es

P. Jodrá: Departamento de Métodos Estadísticos, Universidad de Zaragoza, 50009 Zaragoza, Spain
E-mail address: pjodra@unizar.es