An application of size-bias method

HASSAN HAJI
Department of Statistics
Imam Khomeini International University
Qazvin, IRAN

Abstract: In this paper, we derive an upper bound on the Kolmogorov distance between the distribution of a sum of indicator random variables and a standard normal distribution by using the size-bias method. Also, we give lower and upper bounds for distribution function of sum of indicator random variables in two special points.

Keywords: Indicator random variables, size-biased distribution, Kolmogorov distance.
Received: October 25, 2022. Revised: May 10, 2023. Accepted: June 14, 2023. Published: July 12, 2023.

1. Introduction

Size bias occurs famously in waiting-time paradoxes, undesirably in sampling schemes, and unexpectedly in connection with Stein's method, tightness, analysis of the lognormal distribution, Skorohod embedding, infinite divisibility, and number theory [1,4]. For a non-negative random variable X with $\mu = \mathbf{E}(X) < \infty$, we say a random variable X^s has the size-biased distribution with respect to X if

$$\mathbf{E}(Xf(X)) = \mu\mathbf{E}(f(X^s)),$$

For all $f:[0,\infty)\to \mathrm{R}$ such that $\mathbf{E}|Xf(X)|<\infty$ [2].

Let $Y=\sum_{i=1}^{n} Y_i$, where $Y_i \geq 0$ and

1. Y_i^s have the size-biased distribution of Y_i independent of $(Y_j)_{j\neq i}$ and $(Y_j^s)_{j\neq i}$ for $i=1,\dots,n$.

2. Define a vector $(Y_j^{(i)})_{j\neq i}$ such that its conditional distribution given Y_i^s coincides with that of $(Y_j)_{j\neq i}$ given Y_i.

3. Choose an index J such that

$$P(J=j)=\frac{\mathbf{E}(Y_j)}{\mathbf{E}(Y)}.$$

Then $Y^s=\sum_{k\neq J} Y_k^{(J)} + Y_J^s$ has the size-biased distribution with respect to Y (see Section 2.4 in [2]). For an indicator random variable I, $P(I^s=1)=\frac{P(I=1)}{\mathbf{E}(I)}=1$, which means $I^s=1$. Then in this case, $Y^s=\sum_{k\neq J} Y_k^{(J)}+1$.

The paper is organized as follows. In Section 2, we show the simple calculations related to the sum of indicator random variables on a random permutation. Section 3 is devoted to the proofs of our results. We use the size-bias method to prove Theorem 2 and get an upper bound on the Kolmogorov distance between the distribution of sum of indicator random variables and a standard normal distribution. Also, we give lower and upper bounds for distribution function of Σ_n in two special points. To emphasize the practical usefulness of our results, we note that Σ_n is related to the number of leaves in tree structures. In the other words, the expectation and variance of Σ_n is important for studying of random trees.

2. Preliminaries

Set

$$\mathrm{I}(A) := \begin{cases} 1, & \text{if } A \text{ is true} \\ 0, & \text{otherwise.} \end{cases}$$

Let $t = (t_1, ..., t_n)$ be a permutation on $\{1, 2, ..., n\}$ and

$$\Sigma_n = \sum_{i=1}^{n-1} I_{i,i+1}, \ (n > 1)$$

where $I_{i,j} = \mathrm{I}(t_i > t_j)$. We have the following facts:

$$P(I_{i,i+1} = 1) = \frac{1}{2}, \qquad (1)$$

$$P(I_{i,i+1} I_{j,j+1} = 1) = \begin{cases} \frac{1}{6}, & |i-j| = 1 \\ \frac{1}{4}, & |i-j| > 1. \end{cases} \qquad (2)$$

Let $I_{i,j,k} = \mathrm{I}(t_i > t_j > t_k)$. Then

$$P(I_{i,i+2,i+1} I_{j,j+2,j+1} = 1) \begin{cases} = 0, & |i-j| = 1 \\ \leq \frac{1}{36}, & |i-j| > 1. \end{cases} \qquad (3)$$

Thus from (1),

$$\mathbf{E}(\Sigma_n) = \frac{n-1}{2}.$$

From (2),

$$\mathbf{E}(\Sigma_n^2) = \mathbf{E}(\sum_{i=1}^{n-1} I_{i,i+1} + \sum_{i \neq j} I_{i,i+1} I_{j,j+1})$$

$$= \frac{n-1}{2} + \frac{(n-2)(n-1)}{4} + \frac{2(n-2)}{6}$$

$$= \frac{3n^2 - 5n + 4}{12} \qquad (4)$$

and thus

$$\mathbf{Var}(\Sigma_n) = \frac{n+1}{12}.$$

Since $\mathrm{I}(A^c) = 1 - \mathrm{I}(A)$,

$$\mathbf{Var}(\sum_{i=1}^{n-1} I_{i,i+1}^c) = \mathbf{Var}(n - 1 - \Sigma_n) = \frac{n+1}{12} \qquad (5)$$

and from (3),

$$\mathbf{E}((\sum_{i=1}^{n-2} I_{i,i+2,i+1}^c)^2) \leq \frac{n-2}{6} + \frac{(n-3)(n-4)}{36}$$

$$= \frac{n^2 - n}{36}.$$

Hence

$$\mathbf{Var}(\sum_{i=1}^{n-1} I_{i,i+2,i+1}^c) \leq \frac{3n-4}{36}. \qquad (6)$$

In the same manner,

$$\mathbf{Var}(\sum_{i=2}^{n-1} I_{i,i-1,i+1}^c) \leq \frac{3n-4}{36}. \qquad (7)$$

Theorem 1 *[4] Let X be a nonnegative random variable with mean and variance μ and σ^2, respectively, both finite and positive. Suppose X^s has the size-biased distribution with respect to X which satisfies $|X^s - X| \leq C$ for some constant $C > 0$ with probability one. Let $A = C\mu / \sigma^2$.*

If $X^s \geq X$ with probability one, then

$$F_X(\mu - t\sigma) \leq \exp\left(-\frac{t^2}{2A}\right), \quad \text{for all } t > 0.$$

If the moment generating function $m(\theta) = \mathbf{E}(e^{\theta X})$ is finite at $\theta = 2/C$, then

$$F_X(\mu + t\sigma) \geq 1 - \exp\left(-\frac{t^2}{2(A + Bt)}\right),$$

$$\text{for all } t > 0, \text{where } B = C/2\sigma.$$

Such concentration of measure results are applied to a number of new examples: the number of relatively ordered subsequences of a random permutation, sliding window statistics including the number of m-runs in a sequence of coin tosses, the number of local maxima of a random function on a lattice, the number of urns containing exactly one ball in an urn allocation model, and the volume covered by the union of n balls placed uniformly over a volume n subset of $\mathbb{R}^d$.

3. Main Results

In this section, an upper bound on the Kolmogorov distance between the distribution of a sum of indicator random variables and a standard normal distribution is obtained by using the size-bias method. Also, the lower and upper bounds for distribution function of sum of indicator random variables in two special points is given.

The Wasserstein distance between any two probability measures μ and υ on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ is defined as follows

$$dis^W(\mu,\upsilon)=\sup_{h\in H}\left|\int_{\mathbb{R}}h(x)d\mu(x)-\int_{\mathbb{R}}h(x)d\upsilon(x)\right|,$$

where $H:=\{h:\mathbb{R}\to\mathbb{R}:|h(x)-h(y)|\le|x-y|\}$.

For random variables X and Y, the Kolmogorov distance between their distributions is defined as

$$dis^K(X,Y)=\sup_x|F_X(x)-F_Y(x)|.$$

Also, for a random variable X with Lebesgue density bounded C [7],

$$dis^K(X,Y)\le\sqrt{2C\,dis^W(X,Y)}. \qquad (8)$$

Let X be a non-negative random variable with $\mathbf{E}(X)<\infty$. Let X^s have the size-biased distribution with respect to X. If $T=\frac{X-\mathbf{E}(X)}{\sqrt{\mathbf{Var}(X)}}$ and $Z\approx N(0,1)$, then [6,7]:

$$dis^W(T,Z)\le\frac{\mathbf{E}(X)}{\mathbf{Var}(X)}\sqrt{\frac{2}{\pi}\mathbf{Var}(\mathbf{E}(X^s-X\mid X))}$$

$$+\frac{\mathbf{E}(X)}{\mathbf{Var}(X)^{\frac{3}{2}}}\mathbf{E}((X^s-X)^2). \qquad (9)$$

Using Jensen's inequality for $f(x)=x^2$,

$$\mathbf{Var}(\mathbf{E}(X\mid\mathcal{G}_1))\le\mathbf{Var}(\mathbf{E}(X\mid\mathcal{G}_2)), \qquad (10)$$

where $\mathcal{G}_1,\mathcal{G}_2$ are two sigma-fields, satisfying $\mathcal{G}_1\subseteq\mathcal{G}_2$ [3]. Thus, if $\mathcal{F}=\sigma(I_{1,2},...,I_{n-1,n})$, then $\sigma(\Sigma_n)\subseteq\mathcal{F}$.

Theorem 2 *Suppose* $Z\approx N(0,1)$ *and*

$$T=\frac{\Sigma_n-\frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}}.$$

Then

$$\mathrm{dis}^K(T,Z)\le\left(\sqrt{\frac{2}{\pi}}\left(3\sqrt{3}\frac{\sqrt{n}}{n+1}+12\sqrt{3}\frac{n-1}{\sqrt{(n+1)^3}}\right)\right)^{\frac{1}{2}}.$$

Proof. Choose an index J uniformly at random from the set $\{1,...,n-1\}$, then size-bias $I_{J,J+1}$ by letting it equal to one, and take the remaining summands conditional on $I_{J,J+1}=1$. We can realize $I_{J,J+1}=1$ by adjusting the order of t_J and t_{J+1} such that $t_J>t_{J+1}$, and Σ_n^s denotes the number of descents in t after adjusting the order of t_J and t_{J+1}. Then for $J=1$,

$$M_1:=\Sigma_n^s-\Sigma_n=(I_{1,3}+1-I_{2,3})I_{1,2}^c$$
$$=I_{1,2}^c-I_{1,3}^c,$$

for $J=n-1$,

$$M_{n-1}:=\Sigma_n^s-\Sigma_n=(I_{n-2,n}+1-I_{n-2,n-1})I_{n-1,n}^c$$
$$=I_{n-1,n}^c-I_{n-1,n-2,n}^c,$$

and for $2\le J\le n-2$,

$$M_J:=\Sigma_n^s-\Sigma_n=(I_{J-1,J+1}+1+I_{J,J+2}-I_{J-1,J}-I_{J+1,J+2})I_{J,J+1}^c$$

$$=I_{J,J+1}^c-I_{J,J-1,J+1}^c-I_{J,J+2,J+1}^c.$$

From (5), (6) and (7),

$$\mathbf{Var}(\mathbf{E}(\Sigma_n^s-\Sigma_n\mid\mathcal{F}))=\frac{1}{(n-1)^2}\mathbf{Var}(M_1+\sum_{i=2}^{n-2}M_i+M_{n-1})$$

$$=\frac{1}{(n-1)^2}\mathbf{Var}(\sum_{i=1}^{n-1}I_{i,i+1}^c+\sum_{i=1}^{n-1}I_{i,i+2,i+1}^c+\sum_{i=2}^{n-1}I_{i,i-1,i+1}^c)$$

$$\le\frac{3}{(n-1)^2}(\mathbf{Var}(\sum_{i=1}^{n-1}I_{i,i+1}^c)+\mathbf{Var}(\sum_{i=1}^{n-1}I_{i,i+2,i+1}^c)$$

$$+\mathbf{Var}(\sum_{i=2}^{n-1} I^{c}_{i,i-1,i+1}))$$

$$\leq \frac{3}{(n-1)^2}(\frac{n+1}{12}+2\frac{3n-4}{36})$$
$$=\frac{9n-5}{12(n-1)^2}.$$

Also,

$$\mathbf{E}((\Sigma_n^s-\Sigma_n)^2)=\mathbf{E}(\mathbf{E}((\Sigma_n^s-\Sigma_n)^2 \mid \mathsf{F}))$$
$$=\frac{1}{(n-1)}\mathbf{E}(M_1^2+\sum_{i=2}^{n-2} M_i^2+M_{n-1}^2)$$
$$\leq 1.$$

Proof is completed from (9) and then (8), since $\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \leq \frac{1}{\sqrt{2\pi}}$ for all $x \in \mathsf{R}$.

Suppose $Z : N(0,1)$. It is obvious that for $0 < z < 1$,

$$F_Z(\frac{(n-1)(z-1)}{2\sqrt{(n+1)/12}}) \to 0, \quad n \to \infty$$

and for $z > 1$,

$$F_Z(\frac{(n-1)(z-1)}{2\sqrt{(n+1)/12}}) \to 1, \quad n \to \infty.$$

Theorem 3 *We have*

$$\lim_{n\to\infty} F_{\Sigma_n}(\frac{n-1}{2}s)=\begin{cases}1, & s>1\\ 0, & s<1.\end{cases}$$

Proof . Since $\Sigma_n > 0$, then $F_{\Sigma_n}(\frac{n-1}{2}s)=0$ for $s \leq 0$. Also

$$F_{\Sigma_n}(\frac{n-1}{2}s)=F_T(\frac{(n-1)(s-1)}{2\sqrt{(n+1)/12}}).$$

From Theorem 2 and the definition of Kolmogorov distance,

$$F_T(\frac{(n-1)(s-1)}{2\sqrt{(n+1)/12}}) \leq F_Z(\frac{(n-1)(s-1)}{2\sqrt{(n+1)/12}})+O\left(\frac{1}{n^{\frac{1}{4}}}\right).$$

From (11) and (12), the proof is completed.

Theorem 4 *For* $s > 0$,

$$F_{\Sigma_n}((n-1)(s+1/2)) \geq 1-\exp(-\frac{(n-1)s^2}{1+s})$$

and

$$F_{\Sigma_n}((n-1)(s-1/2)) \leq \exp(-(n-1)s^2).$$

Proof. The inequalities are proved with selection $t=\frac{(n-1)s}{\sqrt{(n+1)/12}}$ in Theorem 1, since $|\Sigma_n-\Sigma_n^s| \leq 1$.

References

[1] Arratia, R. Goldstein, L., Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent, Available in http://bcf.usc.edu/ larry/papers/pdf/csb.pdf, . 2009.
[2] Arratia, A. Goldstein, L. and Kochman, F., (11) Size-bias for one and all, Preprint. Available at arXiv: 1308.2729, 2013.
[3] Billingsley, P., *Probability and Measure*, John Wiley and Sons, New York, 1885. (12)
[4] Ghosh, S. and Goldstein, L., Concentration of measures via size-biased couplings. *Porbability Theory and Related Fields*, 2011, **149**, 271-278.
[5] Ghosh, S. and Goldstein, L., Applications of size biased couplings for concentration of measures, *Electronic Communications in Probability*, 2011, **16**, 70-83.
[6] Goldstein, L. and Rinott, Y., Multivariate normal approximations by Stein's method and size bias couplings, *Journal of Applied Probability*, 1996, **33(1)**, 1-17.
[7] Ross, N., Fundamentals of Stein's method, *Probability Surveys*, 2011, **8**, 210-293.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)
The author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflict of Interest
The author has no conflict of interest to declare that is relevant to the content of this article.