Statistics and Probability Letters 83 (2013) 1254–1259

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A sharp estimate of the binomial mean absolute deviation with applications

Daniel Berend [a,b], Aryeh Kontorovich [a,*]

[a] Department of Computer Science, Ben-Gurion University, Beer Sheva, 84105, Israel

[b] Department of Mathematics, Ben-Gurion University, Beer Sheva, 84105, Israel

ARTICLE INFO

Article history:
Received 1 October 2012
Received in revised form 17 January 2013
Accepted 17 January 2013
Available online 23 January 2013

Keywords:
Binomial
Mean absolute deviation
Density estimation
Total variation

ABSTRACT

We give simple, sharp non-asymptotic bounds on the mean absolute deviation (MAD) of a Bin(n, p) random variable. Although MAD is known to behave asymptotically as the standard deviation, the convergence is not uniform over the range of p and fails at the endpoints. Our estimates hold for all $p \in [0, 1]$ and illustrate a simple transition from the "linear" regime near the endpoints to the "square root" regime elsewhere. As an application, we provide asymptotically optimal tail estimates of the total variation distance between the empirical and the true distributions over countable sets.

1. Introduction

The *mean absolute deviation* (MAD) of a random variable Y is defined by $\mathbb{E}\,|Y - \mathbb{E}Y|$. For $Y \sim \text{Bin}(n, p)$, a closed-form expression for MAD is known, apparently having been first discovered by De Moivre:

$$D_n(p) := 2(1-p)^{n-\lfloor np\rfloor} p^{\lfloor np\rfloor+1} (\lfloor np\rfloor + 1) \binom{n}{\lfloor np\rfloor + 1}, \tag{1}$$

where the floor function is defined by $\lfloor x\rfloor = \max\{n \in \mathbb{N} : n \le x\}$. Diaconis and Zabell (1991) give a fascinating account of the history of (1) and provide generalizations to other distribution families.

Since MAD is a measure of dispersion of the random variable,[1] it is natural to compare it with more familiar quantities, such as the standard deviation:

$$S_n(p) := \sqrt{\mathbb{E}(Y - \mathbb{E}Y)^2} = \sqrt{np(1-p)}.$$

Blyth (1980) discusses this comparison in some detail, observing the obvious relation

$$D_n(p) \le S_n(p)$$

and also showing that

$$\lim_{n\to\infty} \frac{S_n(p)}{D_n(p)} = \sqrt{\pi/2}, \quad 0 < p < 1. \tag{2}$$

* Corresponding author.
E-mail addresses: berend@cs.bgu.ac.il (D. Berend), karyeh@cs.bgu.ac.il, lkontor@gmail.com (A. Kontorovich).

[1] In some sense, MAD is more natural than standard deviation, but the latter is usually preferred for reasons of analytic tractability.

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2013.01.023

The convergence rate of $S_n(p)/D_n(p)$ may be quantified (Blyth, 1980) by

$$D_n(p) = \sqrt{2/\pi}S_n(p) + O(n^{-1/2}) \quad 0 < p < 1.$$

Similar asymptotics were obtained by Frame (1945) and Johnson (1957). Note, however, that this convergence cannot be uniform in p, since

$$\lim_{p\to 0}\frac{S_n(p)}{D_n(p)} = \lim_{p\to 1}\frac{S_n(p)}{D_n(p)} = \infty, \quad n \in \mathbb{N}. \tag{3}$$

Our main technical contribution is a sharp estimate bound on $D_n(p)$ that holds for all $p \in [0, 1]$.

Theorem 1.

$$2np(1-p)^n = D_n(p) \le 2np, \quad n \in \mathbb{N},\ p < 1/n$$
$$S_n(p)/\sqrt{2} \le D_n(p) \le S_n(p), \quad n \ge 2,\ p \in [1/n, 1-1/n]$$
$$2np^n(1-p) = D_n(p) \le 2n(1-p), \quad n \in \mathbb{N},\ p > 1-1/n.$$

This formula illustrates the behavior alluded to in the Abstract and in particular proves (3). For p near 0 (resp., 1), $D_n(p)$ is roughly linear in np (resp., $n(1-p)$). For p sufficiently far from the endpoints, $S_n(p)$ behaves roughly as $\sqrt{np(1-p)}$. Given (1), the only nontrivial relation of the six stated in Theorem 1 is $S_n(p) \le \sqrt{2}D_n(p)$.[2] The latter is tight (equality is achieved for $n = 2$, $p = 1/2$), but the constant $\sqrt{2}$ may be improved if we restrict our attention to larger n—but obviously not below $\sqrt{\pi/2}$, as per (2). In fact, an inspection of the proof of Theorem 1 yields the following:

Corollary 2.

$$\lim_{n\to\infty}\sup_{p\in[1/n,1-1/n]}\frac{S_n(p)}{D_n(p)} = \frac{e}{2}.$$

It is instructive to compare the above with (2), which holds for a fixed $p \in (0, 1)$.

We apply the estimates in Theorem 1 to obtain asymptotically optimal tail estimates of the total variation distance between the empirical and the true distributions over countable sets.

2. Main results

The relations in Theorem 1 hold for any single binomial variable Y, and we will apply it simultaneously to an infinite ensemble $\{Y_j\}_{j\in\mathbb{N}}$ as follows. Suppose one empirically estimates the distribution $\boldsymbol{p} = (p_1, p_2, \ldots)$ of an $\mathbb{N}$-valued random variable X from n i.i.d. samples X_i via $\hat{\boldsymbol{p}}^{(n)}$, where

$$\hat{p}_j^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i=j\}}, \quad j \in \mathbb{N}.$$

Then, defining

$$J_n := \left\|\hat{\boldsymbol{p}}^{(n)} - \boldsymbol{p}\right\|_1 = \sum_{j\in\mathbb{N}}\left|p_j - \hat{p}_j^{(n)}\right|,$$

we have

$$n\mathbb{E}J_n = \sum_{j\in\mathbb{N}}\mathbb{E}\left|Y_j - np_j\right|, \tag{4}$$

where $Y_j \sim \mathrm{Bin}(n, p_j)$.

For the simple case where $\boldsymbol{p}$ has finite support, we have

$$\mathbb{E}J_n \le \sqrt{\frac{k}{n}}, \qquad \boldsymbol{p} \in \mathbb{R}^k, \quad n \in \mathbb{N}. \tag{5}$$

For $\boldsymbol{p}$ with infinite support, the estimate

$$\mathbb{E}J_n \le \frac{1}{\sqrt{n}}\sum_{j\in\mathbb{N}}\sqrt{p_j} \tag{6}$$

is informative for "most" common distributions; (5) and (6) are proved in Lemma 5.

[2] For $p = 1/2$, note the resemblance to Khintchine's inequality (Szarek, 1976).

The interesting case, however, is $\sum_{j\in\mathbb{N}} \sqrt{p_j} = \infty$, which necessitates invoking the results on MAD for binomial variables. As we show in Lemma 6,

$$\mathbb{E}J_n \leq \alpha_n(\boldsymbol{p}) + \beta_n(\boldsymbol{p}), \tag{7}$$

where

$$\alpha_n(\boldsymbol{p}) = 2\sum_{p_j<1/n} p_j, \qquad \beta_n(\boldsymbol{p}) = \frac{1}{\sqrt{n}} \sum_{p_j\geq 1/n} \sqrt{p_j}. \tag{8}$$

Furthermore,

$$\alpha_n + \beta_n \underset{n\to\infty}{\longrightarrow} 0, \tag{9}$$

although the rate of decay in (9) depends on $\boldsymbol{p}$ and may be arbitrarily slow (Lemmas 7 and 8).

Moreover, our estimate in (7) for $\mathbb{E}J_n$ in terms of α_n and β_n is nearly tight, in the following sense.

Proposition 3. *For all $n \geq 2$ and all distributions $\boldsymbol{p} \in \mathbb{R}^{\mathbb{N}}$,*

$$\mathbb{E}J_n \geq \frac{\alpha_n + \beta_n}{4} - \frac{1}{4\sqrt{n}}.$$

Remark 4. To keep the expressions simple, we have chosen $1/n$ as the break-point in defining α_n and β_n. We note in passing that a minor improvement in the constants is achieved by the (optimal) break-point $1/4n$.

3. Proofs

Proof of Theorem 1. As mentioned above, only the relation

$$S_n(p) \leq \sqrt{2}D_n(p) \tag{10}$$

requires proof. Let us rewrite the mean absolute deviation formula (1) as

$$\mathbb{E}\,|Y - np| = 2k\binom{n}{k} p^k(1-p)^{n-k+1}, \quad (k = \lfloor np\rfloor + 1).$$

Denote the right-hand side by $E(n, k, p)$, and put $G(n, k, p) = 2E(n, k, p)^2/(p(1-p))$. Then (10) is equivalent to the claim

$$G(n, k, p) \geq n, \quad p \in [1/n, 1-1/n], \ (k = \lfloor np\rfloor + 1). \tag{11}$$

The domain where (11) is to be proved may be reparametrized by the inequalities

$$2 \leq k \leq n-1, \quad \frac{k-1}{n} \leq p < \frac{k}{n}.$$

Now the function $G(n, k, \cdot)$ is increasing on $[(k-1)/n, (2k-1)/2n]$ and decreasing on $[(2k-1)/2n, k/n]$—and hence we need only consider the endpoints $p = (k-1)/n$ and $p = k/n$.

To examine the first possibility, we take $p = (k-1)/n$ and seek a k that minimizes $G(n, k, (k-1)/n)$. To this end, we consider the inequality $G(n, k+1, k/n) \geq G(n, k, (k-1)/n)$, which is equivalent (after a routine calculation) to

$$\left(\frac{k}{k-1}\right)^{2k-1} \geq \left(\frac{n-k+1}{n-k}\right)^{2n-2k+1}. \tag{12}$$

Since the function $f(x) = (1+1/x)^{2x+1}$ is monotonically decreasing on $[1, \infty)$, inequality (12) holds whenever $k \leq (n+1)/2$. We conclude that $G(n, k, (k-1)/n)$ is minimized at the smallest allowed value of k, which is $k = 2$. We easily verify that the inequality $G(n, 2, 1/n) \geq n$ is equivalent to $8(n-1)^{2n-1} \geq n^{2n-1}$ for all $n \geq 2$, which again follows from the monotonicity of $(1 + 1/x)^{2x+1}$.

The second case, $p = k/n$, is analyzed in an exactly analogous manner. □

Lemma 5. *Suppose $n \in \mathbb{N}$ and $\boldsymbol{p} \in \mathbb{R}^{\mathbb{N}}$ is a distribution. Then*

$$\sqrt{n}\mathbb{E}J_n \leq \sum_{j\in\mathbb{N}} \sqrt{p_j}.$$

If additionally $\boldsymbol{p} \in \mathbb{R}^k$ has finite support, then

$$\sqrt{n}\mathbb{E}J_n \leq \sqrt{k}.$$

Proof. Let $Y_j \sim \mathrm{Bin}(n, p_j)$. Then

$$\left(\mathbb{E}\left|Y_j - np_j\right|\right)^2 \leq \mathbb{E}(Y_j - np_j)^2 = np_j(1-p_j) \leq np_j,$$

whence

$$\mathbb{E}\left|Y_j - np_j\right| \leq \sqrt{np_j(1-p_j)} \leq \sqrt{np_j}. \tag{13}$$

The first claim follows by (4).

To prove the second claim, define $\boldsymbol{x} \in \mathbb{R}^k$ by $x_j = \sqrt{p_j}$ and recall that

$$\sum_{j=1}^{k} \sqrt{p_j} = \|x\|_1 \leq \sqrt{k}\,\|x\|_2 = \sqrt{k}. \quad \square \tag{14}$$

Lemma 6. *Suppose $n \in \mathbb{N}$ and $\boldsymbol{p} \in \mathbb{R}^{\mathbb{N}}$ is a distribution. Then*

$$\mathbb{E}J_n \leq \alpha_n + \beta_n.$$

Proof. As in the proof of Lemma 5, let $Y_j \sim \mathrm{Bin}(n, p_j)$ and use (4) to write

$$n\mathbb{E}J_n = \sum_{p_j < 1/n} \mathbb{E}\left|Y_j - np_j\right| + \sum_{p_j \geq 1/n} \mathbb{E}\left|Y_j - np_j\right|. \tag{15}$$

The first term on the right-hand side of (15) is upper-bounded by $n\alpha_n(\boldsymbol{p})$ via Theorem 1, while the second term is at most $n\beta_n(\boldsymbol{p})$ via (13). $\square$

Lemma 7. *Let $\boldsymbol{p} \in \mathbb{R}^{\mathbb{N}}$ be a distribution. Then*

$$\alpha_n(\boldsymbol{p}) + \beta_n(\boldsymbol{p}) \underset{n\to\infty}{\longrightarrow} 0.$$

Proof. The decay of $\alpha_n(\boldsymbol{p})$ to zero is obvious, since it is the tail of a convergent series. To prove that

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{p_j \geq 1/n} \sqrt{p_j} = 0, \tag{16}$$

observe that for any $\varepsilon_n > 1/n$,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{p_j \geq 1/n} \sqrt{p_j} &= \sum_{p_j \geq 1/n} \frac{p_j}{\sqrt{np_j}} \\
&= \sum_{1/n \leq p_j \leq \varepsilon_n} \frac{p_j}{\sqrt{np_j}} + \sum_{p_j > \varepsilon_n} \frac{p_j}{\sqrt{np_j}} \\
&\leq \sum_{p_j \leq \varepsilon_n} p_j + \frac{1}{\sqrt{n}} \sum_{p_j > \varepsilon_n} \sqrt{p_j}.
\end{aligned}$$

Choosing $\varepsilon_n = n^{-1/3}$ ensures that both terms in the last expression vanish as $n \to \infty$. $\square$

Lemma 8. *For any rate sequence $1 > r_1 > r_2 > \cdots \searrow 0$, there is a distribution $\boldsymbol{p} \in \mathbb{R}^{\mathbb{N}}$ such that*

$$\alpha_n(\boldsymbol{p}) + \beta_n(\boldsymbol{p}) > r_n, \quad n \in \mathbb{N}.$$

Proof. It suffices to show that there is no rate sequence bounding α_n. But this is obvious, since α_n may be expressed as the tail of a series converging to 2—and although any such tail must decay to zero, the rate may be arbitrarily slow. In fact, given some rate sequence (r_n), to ensure that $\sum_{p_j \geq 1/n} p_j \leq 1 - r_n$ for each $n \in \mathbb{N}$, we may choose the appropriate p_j in an iterative greedy fashion, for $n = 1, 2, \ldots$. $\square$

Remark 9. Orlitsky (2012) points out that this is a special case of a more general phenomenon. For any estimator $\hat{\boldsymbol{q}}^{(n)}$ and any functions $f_n(\boldsymbol{p})$ and $g_n(\varepsilon)$ that decrease to 0 with n such that

$$\mathbb{P}\left[\left\|\hat{\boldsymbol{q}}^{(n)} - \boldsymbol{p}\right\|_1 > f_n(\boldsymbol{p}) + \varepsilon\right] < g_n(\varepsilon),$$

the rate at which $f_n(\boldsymbol{p})$ decreases to 0 must depend on $\boldsymbol{p}$. This is because for any estimator $\hat{\boldsymbol{q}}^{(n)}$ and every n, there is a distribution $\boldsymbol{p}$ such that $\mathbb{P}\left[\left\|\hat{\boldsymbol{q}}^{(n)}-\boldsymbol{p}\right\|_1 > 2-\varepsilon\right] > 1-\varepsilon$. Namely, let $\boldsymbol{p}$ be the uniform distribution over a random $K(n)$-element subset of $\{1,\ldots,N(n)\}$ where $N(n) \gg K(n) \gg n$. Then $\hat{\boldsymbol{q}}^{(n)}$ will have information only about the n elements that appeared and not about the remaining $K(n)-n$ elements, and typical $\hat{\boldsymbol{q}}^{(n)}$ constructed in this way will be very far from $\boldsymbol{p}$. This remark applies in light of (17).

Proof of Proposition 3. Let $n \geq 2$ and $Y_j \sim \mathrm{Bin}(n, p_j)$. We group the probabilities as follows: $P_1 = \left\{j : p_j < 1/n\right\}$, $P_2 = \left\{j : 1/n \leq p_j \leq 1/2\right\}$ and $P_3 = \left\{j : p_j > 1/2\right\}$. By Theorem 1,

$$\mathbb{E}\left|Y_j - np_j\right| \geq \frac{1}{2}\begin{cases} np_j, & j \in P_1, \\ \sqrt{np_j}, & j \in P_2. \end{cases}$$

Now

$$n\alpha_n(\boldsymbol{p}) = \sum_{j\in P_1} 2np_j \leq 4\sum_{j\in P_1}\mathbb{E}\left|Y_j - np_j\right|$$

and

$$\begin{aligned} n\beta_n(\boldsymbol{p}) &= \sum_{j:p_j\geq 1/n}\sqrt{np_j} \\ &\leq \sum_{j\in P_2}\sqrt{np_j} + \sqrt{n} \\ &\leq 2\sum_{j\in P_2}\mathbb{E}\left|Y_j - np_j\right| + \sqrt{n} \end{aligned}$$

and thus

$$4\sum_{j\in P_1}\mathbb{E}\left|Y_j - np_j\right| + 2\sum_{j\in P_2}\mathbb{E}\left|Y_j - np_j\right| + \sqrt{n} \geq n\alpha_n + n\beta_n,$$

which proves the claim. □

Remark 10. An inspection of the proof shows that for large n, the constant 4 in the statement of Proposition 3 may be improved to $e + o(1)$.

4. Application: complete convergence of J_n

Complete convergence was introduced in Hsu and Robbins (1947). Applied to the random variable J_n, it means that

$$\sum_{n=1}^{\infty}\mathbb{P}\left[J_n > \varepsilon\right] < \infty$$

for all $\varepsilon > 0$. Obviously, complete convergence implies almost-sure convergence. The complete convergence of J_n to 0 may be surmised from Sanov's theorem (Cover and Thomas, 2006; den Hollander, 2000)—whose drawback, however, is that it does not readily yield explicit, analytically tractable estimates for $\mathbb{P}\left[J_n > \varepsilon\right]$.

As noted in Pinelis (1990) and Devroye (1991), such estimates on J_n are readily available via McDiarmid's inequality (McDiarmid, 1989), which yields

$$\mathbb{P}\left[|J_n - \mathbb{E}J_n| > \varepsilon\right] \leq 2\exp(-n\varepsilon^2/2), \quad n \in \mathbb{N},\ \varepsilon > 0. \tag{17}$$

In order for (17) to be effectively applicable, one must be able to bound $\mathbb{E}J_n$. That was, in fact, the initial motivation behind this paper, whose main result may be summarized as

$$\frac{1}{4}(\alpha_n + \beta_n - n^{-1/2}) \leq \mathbb{E}J_n \leq \beta_n + \min\left\{\alpha_n, \frac{1}{\sqrt{n}}\sum_{p_j<1/n}\sqrt{p_j}\right\}, \quad n \geq 2.$$

Besides having asymptotically tight bounds on $\mathbb{E}J_n$, we also remark that the coefficient $1/2$ in the exponent in (17) cannot, in general, be improved. This follows from Sanov's theorem and a reverse Pinsker inequality (Berend et al., 2012).

Acknowledgments

We thank Andrew Barron, László Györfi, John Hartigan, Gábor Lugosi, David McAllester, Alon Orlitsky, David Pollard, and Larry Wasserman for helpful discussions and comments, and an anonymous referee for corrections. We are grateful to the Stone family for providing a venue for this work. The second author was supported in part by the Israel Science Foundation (grant No. 1141/12).

References

Berend, Daniel, Harremoës, Peter, Kontorovich, Aryeh, 2012. A reverse Pinsker inequality. Preprint.
Blyth, Colin R., 1980. Expected absolute error of the usual estimator of the binomial parameter. Amer. Statist. 34 (3), 155–157.
Cover, Thomas M., Thomas, Joy A., 2006. Elements of Information Theory, second ed. Wiley-Interscience, Hoboken, NJ.
den Hollander, Frank, 2000. Large Deviations. In: Fields Institute Monographs, vol. 14. American Mathematical Society, Providence, RI.
Devroye, Luc, 1991. Exponential inequalities in nonparametric estimation. In: Nonparametric Functional Estimation and Related Topics (Spetses, 1990). In: NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 335. Kluwer Acad. Publ., Dordrecht, pp. 31–44.
Diaconis, Persi, Zabell, Sandy, 1991. Closed form summation for classical distributions: variations on a theme of de Moivre. Statist. Sci. 6 (3), 284–302.
Frame, James S., 1945. Mean deviation of the binomial distribution. Amer. Math. Monthly 52 (7), 377–379.
Hsu, Pao-Lu, Robbins, Herbert, 1947. Complete convergence and the law of large numbers. Proc. Natl. Acad. Sci. USA 33, 25–31.
Johnson, Norman L., 1957. A note on the mean deviation of the binomial distribution. Biometrika 44 (3–4), 532–533.
McDiarmid, Colin, 1989. On the method of bounded differences. In: Siemons, J. (Ed.), Surveys in Combinatorics. In: LMS Lecture Notes Series, vol. 141. Morgan Kaufmann Publishers, San Mateo, CA, pp. 148–188.
Orlitsky, Alon, 2012. Convergence of the empirical distribution. Private communication.
Pinelis, Iosif F., 1990. Inequalities for distributions of sums of independent random vectors and their application to estimating a density. Teor. Veroyatn. Primen. 35 (3), 592–594.
Szarek, Stanisław J., 1976. On the best constants in the Khinchin inequality. Studia Math. 58 (2), 197–208.