A SHORT NOTE ON THE AVERAGE MAXIMAL NUMBER OF BALLS IN A BIN

MARCUS MICHELEN

ABSTRACT. We analyze the asymptotic behavior of the average maximal number of balls in a bin obtained by throwing uniformly at random r balls without replacement into n bins, T times. Writing the expected maximum as $\frac{r}{n}T + C_{n,r}\sqrt{T} + o(\sqrt{T})$, a recent preprint of Behrouzi-Far and Zeilberger asks for an explicit expression for $C_{n,r}$ in terms of n, r and π. In this short note, we find an expression for $C_{n,r}$ in terms of n, r and the expected maximum of n independent standard Gaussians. This provides asymptotics for large n as well as closed forms for small n—e.g. $C_{4,2} = \frac{3}{2\pi^{3/2}} \arccos(-1/3)$—and shows that computing a closed form for $C_{n,r}$ is precisely as hard as the difficult question of finding the expected maximum of n independent standard Gaussians.

1. Introduction

Suppose that you have n bins, and in each round, you throw r balls such that each ball lands in a different bin, with each of the $\binom{n}{r}$ possibilities equally likely. After T rounds, set $U(n, r; T)$ to be the maximum occupancy among the n bins. Set $A(n, r; T) = \mathbb{E}U(n, r; T) - \frac{r}{n}T$, and suppose that $A(n, r; T) = C_{n,r}\sqrt{T} + o(\sqrt{T})$. A recent preprint [2] of Behrouzi-Far and Zeilberger asks for an explicit expression for $C_{n,r}$ in terms of n, r and π; they also calculate estimates for $C_{2,1}, C_{3,1}, C_{4,1}, C_{4,2}$ using recurrence relations derived with computer aid. As a motivation, Behrouzi-Far and Zeilberger [2] note that this problem arises in computer systems since load distribution across servers can be modeled with balls and bins.

Rather than utilizing exact computation in the vein of [2], we use a multivariate central limit theorem to prove the following:

Theorem 1.1.

$$C_{n,r} := \lim_{T\to\infty} \frac{A(n, r; T)}{\sqrt{T}} = \sqrt{\frac{r(n-r)}{n(n-1)}}\,\mathbb{E}\left[\max_{1\leqslant j\leqslant n} Z_j\right]$$

where Z_j are i.i.d. standard Gaussians.

The expected maximum of n i.i.d. standard Gaussians appears to have no known closed form for general n and in fact the known forms for small n can be quite nasty; for instance, when $n = 5$, the expected value is $\frac{5}{2\pi^{3/2}} \arccos(-23/27)$. A short table of computed values is included in Section 3.

From here, we extract asymptotics for $n \to \infty$, uniformly in r:

Corollary 1.2. *As $n \to \infty$, we have*

$$C_{n,r} \sim \sqrt{\frac{2r(n-r)\log(n)}{n^2}}$$

uniformly in r.

Proof. This follows from utilizing $\mathbb{E}[\max_{1\leqslant j\leqslant n} Z_j] \sim \sqrt{2\log(n)}$ (see, for instance, [3, Exercise 3.2.3]). ☐

The exact form in Theorem 1.1 also picks up a nice combinatorial property:

Corollary 1.3. *For each n, the sequence $\{C_{n,r}\}_{r=1}^{n-1}$ is* log-*concave.*

Proof. Log-concavity follows from the inequality

$$(r-1)(n-r+1)(r+1)(n-r-1) = (r^2-1)((n-r)^2-1) \leqslant r^2(n-r)^2\,.$$

☐

To prove Theorem 1.1, we use a multivariate central limit theorem to prove a limit theorem for $\frac{U(n,r;T)-\frac{r}{n}T}{\sqrt{T}}$ (Corollary 2.2), show that we can exchange the limit and expectation (Lemma 2.3), and then relate this expectation to the expected maximum of i.i.d. standard normals (Lemma 2.4).

2. Proving Theorem 1.1

Set $b(n,r;T)$ to be the random vector in $\{0,1,\dots,T\}^n$ denoting the occupancies of the bins at time T. The following representation for $b(n,r;T)$ is immediate:

Lemma 2.1. *Fix n,r and let X be the random variable in $\{0,1\}^n$ chosen uniformly among vectors $v\in\{0,1\}^n$ with $\|v\|_{L^1}=r$. Let $X_1,X_2,\dots$ be i.i.d. copies of X. Then $b(n,r;T)\overset{d}{=}\sum_{j=1}^T X_j$. Further, the random variable X has covariance matrix Γ given by*

$$\Gamma_{i,j} = \begin{cases} \frac{r(n-r)}{n^2} & \text{for } i=j \\ -\frac{r(n-r)}{n^2(n-1)} & \text{for } i\neq j \end{cases}\,.$$

Proof. The covariance matrix Γ can be calculated easily:

$$\Gamma_{j,j} = \frac{r}{n}\left(1-\frac{r}{n}\right) = \frac{r(n-r)}{n^2}\,.$$

For $\Gamma_{i,j}$ with $i\neq j$, we compute

$$\Gamma_{i,j} = \frac{\binom{n-2}{r-2}}{\binom{n}{r}} - \frac{r^2}{n^2} = -\frac{r(n-r)}{n^2(n-1)}\,.$$

☐

From here, the multivariate central limit theorem shows convergence in distribution.

Corollary 2.2.

$$\frac{U(n,r;T)-\frac{r}{n}T}{\sqrt{T}} \xrightarrow{d} \max\{Y_1,\dots,Y_n\}$$

where $(Y_1,\dots,Y_n)$ is a mean-zero multivariate Gaussian with covariance matrix Γ, given in Lemma 2.1.

Proof. The multivariate central limit theorem [3, Theorem 3.9.6] implies that

$$\frac{b(n,r;T)-\mathbb{E}b(n,r;T)}{\sqrt{T}} \to \mathcal{N}(0,\Gamma)\,.$$

The identity $\mathbb{E}b(n,r;T) = (\frac{r}{n},\dots,\frac{r}{n})$ together with the continuous mapping theorem implies the Corollary. ☐

To gain information about $A(n,r;T)$, we need to show that not only do we have convergence in distribution, but that we can switch the order of taking limits and expectation.

Lemma 2.3.

$$C_{n,r} := \lim_{T\to\infty} \frac{A(n,r;T)}{\sqrt{T}} = \mathbb{E}\max\{Y_1,\ldots,Y_n\}$$

where $(Y_1,\ldots,Y_n)$ are jointly Gaussian with mean 0 and covariance matrix given by Γ as defined in Lemma 2.1.

Proof. Our strategy is to show uniform integrability of $\widehat{U}(T) := (U(n,r;T) - \frac{r}{n}T)/\sqrt{T}$; for $j \in \{1,2,\ldots,n\}$, let $b^{(j)}$ denote the number of balls in bin j. Then by a union bound, we have

$$\mathbb{P}\left[\left|U(n,r;T) - \frac{r}{n}T\right| \geqslant \lambda\sqrt{T}\right] \leqslant n\mathbb{P}\left[\left|b^{(1)} - \frac{r}{n}T\right| \geqslant \lambda\sqrt{T}\right]. \tag{1}$$

By Hoeffding's inequality (e.g. [1, Theorem 7.2.1]), we bound

$$\mathbb{P}\left[\left|b^{(1)} - \frac{r}{n}T\right| \geqslant \lambda\sqrt{T}\right] \leqslant 2\exp\left(-2\lambda^2\right).$$

Thus, for each T and $K > 0$ we have

$$\mathbb{E}\left[|\widehat{U}(T)| \cdot \mathbf{1}_{|\widehat{U}(T)|\geqslant K}\right] \leqslant 2n\int_K^\infty e^{-2\lambda^2}\, d\lambda\,.$$

This goes to zero uniformly in T as $K \to \infty$, thereby showing that the family $\{\widehat{U}(T)\}_{T\geqslant 0}$ is uniformly integrable. Since uniform integrability together with convergence in distribution implies convergence of means, Corollary 2.2 completes the proof. ∎

All that remains now is to relate $\mathbb{E}\max\{Y_1,\ldots,Y_n\}$ to the right-hand-side of Theorem 1.1.

Lemma 2.4. *Let $(Y_1,\ldots,Y_n)$ be jointly Gaussian with mean 0 and covariance matrix Γ. Then*

$$\mathbb{E}[\max\{Y_1,\ldots,Y_n\}] = \sqrt{\frac{r(n-r)}{n(n-1)}}\,\mathbb{E}\left[\max_{1\leqslant j\leqslant n} Z_j\right]$$

where the variables Z_j are i.i.d. standard Gaussians.

Proof. Consider a multivariate Gaussian $(W_1,\ldots,W_n)$ with mean 0 and covariance matrix given by

$$\widetilde{\Gamma}_{i,j} = \begin{cases} \frac{n}{n-1} & \text{for } i = j \\ -\frac{n}{(n-1)^2} & \text{for } i \neq j\,. \end{cases}$$

Since $\Gamma = \frac{r(n-r)(n-1)}{n^3}\widetilde{\Gamma}$, we have

$$(Y_1,\ldots,Y_n) \overset{d}{=} \sqrt{\frac{r(n-r)(n-1)}{n^3}}(W_1,\ldots,W_n)\,. \tag{2}$$

The vector $(W_1,\ldots,W_n)$ can in fact be realized by setting $W_j = Z_j - \frac{\sum_{i\neq j} Z_i}{n-1}$ with Z_i i.i.d. standard Gaussians. This is because the two vectors are both mean-zero multivariate Gaussians and have the same covariance matrix. Setting $S_n = \sum_{i=1}^n Z_i$, we note

$$W_j = -\frac{S_n}{n-1} + \frac{n}{n-1}Z_j$$

thereby implying

$$\max_{1\leqslant j\leqslant n}\{W_j\} = -\frac{S_n}{n-1} + \left(\frac{n}{n-1}\right)\max_{1\leqslant j\leqslant n}\{Z_j\}\,.$$

Taking expectations and utilizing (2) completes the proof. □

Remark 2.5. The final piece of the proof of Lemma 2.4—relating the expected maximum of the process $(Z_j - \frac{\sum_{i\neq j} Z_i}{n-1})_{j=1}^n$ to that of $(Z_j)_{j=1}^n$—is due to a Math Overflow answer of Iosef Pinelis [4].

Proof of Theorem 1.1: The theorem follows by combining Lemmas 2.3 and 2.4. □

3. Comparison with Numerics

Theorem 1.1 proves an equality for $C_{n,r}$, although for large n, the expectation on the right-hand-side of Theorem 1.1 appears to have no known closed form. Calculating these values for small n is tricky and tedious; we reproduce a few values of $\mathbb{E}[\max_{1\leqslant j\leqslant n} Z_j]$ which can be computed precisely, as calculated in [5]:

n	$\mathbb{E}[\max_{1\leqslant j\leqslant n} Z_j]$
2	$\pi^{-1/2}$
3	$(3/2)\pi^{-1/2}$
4	$3\pi^{-3/2}\arccos(-1/3)$
5	$(5/2)\pi^{-3/2}\arccos(-23/27)$

We can then use these to obtain exact values for the values of $C_{n,r}$ predicted in [2], and note that their predictions are quite close:

	Exact Value	Numerical Approximation	Predicted Value from [2]
$C_{2,1}$	$\frac{1}{\sqrt{2\pi}}$	0.39894...	0.3989...
$C_{3,1}$	$\frac{\sqrt{3}}{2\sqrt{\pi}}$	0.48860...	0.489...
$C_{4,1}$	$\frac{3}{2\pi^{3/2}}\arccos(-1/3)$	0.51469...	0.516...
$C_{4,2}$	$\frac{\sqrt{3}}{\pi^{3/2}}\arccos(-1/3)$	0.59431...	0.59430...

References

[1] N. Alon and J. H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., Hoboken, NJ, third edition, 2008. With an appendix on the life and work of Paul Erdős.

[2] A. Behrouzi-Far and D. Zeilberger. On the average maximal number of balls in a bin resulting from throwing r balls into n bins t times. *arXiv preprint arXiv:1905.07827*, 2019.

[3] R. Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition, 2010.

[4] I. Pinelis. Expectation of maximum of multivariate gaussian. MathOverflow. URL:https://mathoverflow.net/q/332113 (version: 2019-05-21).

[5] A. Selby. Expected value for maximum of a normal random variable. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/510580 (version: 2013-10-01).

Dept. of Mathematics, University of Pennsylvania, 209 South 33rd Street, Philadelphia, PA 19104.

E-mail address: `marcusmi@sas.upenn.edu`