BEYOND THE PIGEON-HOLE PRINCIPLE: MANY PIGEONS IN THE SAME BOX

STEVEN J. MILLER

ABSTRACT. Consider N boxes and m balls, with each ball equally likely to be in each box. For fixed k, we bound the probability of at least k balls being in the same box, as N and m tend to infinity. In particular, we show that if $m = N^{\frac{k-1}{k}}$ then this probability is at least $\frac{1}{k!} - \frac{1}{2\cdot k!^2} + O\left(N^{-1/k}\right)$ and at most $\frac{1}{k!} + O\left(N^{-1/k}\right)$. We then investigate what happens when k grows with N and m, and show there is negligible probability of having at least N balls in the same box when $m = N^{2-\epsilon}$.

1. INTRODUCTION

Dirichlet's Pigeon Hole Principle states that if $N+1$ balls are placed in N boxes, then at least one box must contain at least two balls. We can instead ask how many balls we need (as a function of N) to ensure a 50% (or at least a positive percent independent of N) chance that one box has two balls. This is the classic birthday problem; the probability that $m \le N$ balls are placed in m different boxes is just

$$P_{N,m} \ = \ \frac{N}{N} \cdot \frac{N-1}{N} \cdots \frac{N-(m-1)}{N} \ = \ \frac{N!}{(N-m)!N^m}. \tag{1.1}$$

Hence the probability that at least one box has at least two balls is $1 - P_{N,m}$. To obtain a positive percent we need $P_{N,m}$ bounded away from 1; this occurs when $m \sim \sqrt{N}$. One way to see this is to use Stirling's formula, which says

$$n! \ = \ n^n e^{-n} \sqrt{2\pi n} \left(1 + O(n^{-1})\right), \tag{1.2}$$

as well as

$$e^x \ = \ \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n. \tag{1.3}$$

Thus we find

$$\begin{aligned} P_{N,m} \quad &\sim \quad \frac{N^N e^{-N}\sqrt{2\pi N}}{(N-m)^{N-m} e^{-(N-m)} \sqrt{2\pi(N-m)} \cdot N^m} \\ &\sim \quad \sqrt{\frac{N}{N-m}} \ \left(1 - \frac{m}{N}\right)^{-(N-m)} \ e^{-m} \end{aligned} \tag{1.4}$$

the above is bounded away from 1 when $m \sim \sqrt{N}$.

We consider the more general situation, namely, how many balls are needed to ensure a positive probability of having at least k balls in a box. Here we consider k fixed and $1 \ll m \ll N$, with m and N tending to infinity.

Let $|E|$ denote the probability of an event E, and let $E_{k;N,m}$ be the event that at least k of the m balls are in one of the N boxes. Our main result is

Theorem 1.1. *Let k be fixed. If $m = N^{\frac{k-1}{k}}$, then as $N \to \infty$ we have*

$$\frac{2 \cdot k! - 1}{2 \cdot k!^2} + O\left(N^{-1/k}\right) \ < \ |E_{k;N,m}| \ < \ \frac{1}{k!} + O\left(N^{-1/k}\right). \tag{1.5}$$

2. Proof of Theorem 1.1

We first establish some notation before proving Theorem 1.1. We fix a pair (N, m) with $1 \ll m \ll N$; N and m will tend to infinity. Let $E_{k,i;N,m}$ denote the event of at least k balls in box i (with N boxes and m balls), and let $E_{k;N,m}$ denote the event of at least k balls in a box (with N boxes and m balls). Clearly

$$E_{k;N,m} \ = \ \bigcup_{i=1}^{N} E_{k,i;N,m}. \tag{2.6}$$

However, for N and m even modestly sized, the events $E_{k,1;N,m}, \dots, E_{k,N;N,m}$ are not independent. Thus we obtain an upper bound for the probability of at least k of the m balls in one of the N boxes:

$$|E_{k;N,m}| \ < \ \sum_{i=1}^{N} |E_{k,i;N,m}| \ = \ N \cdot |E_{k,1;N,m}|, \tag{2.7}$$

where the last follows from symmetry.

Proof of the Upper Bound in (1.5). Let $F_{n,1;N,m}$ denote the event of *exactly* n of the m balls being in the first of the N boxes. Then

$$|F_{n,1;N,m}| \quad = \quad \binom{m}{n} \frac{1}{N^n} \binom{m-n}{m-n} \left(1 - \frac{1}{N}\right)^{m-n}. \tag{2.8}$$

We first analyze the case when $n = k$, the main term. For $m \gg k$, we find

$$|F_{k,1;N,m}| \ = \ \frac{1}{k!} \frac{m^k}{N^k}\, e^{-m/N} + O(N^{-1}). \tag{2.9}$$

For all n we can bound $|F_{n,1;N,m}|$ by $\frac{1}{n!}\frac{m^n}{N^n}$, and thus

$$\begin{aligned} |E_{k,1;N,m}| \quad &= \quad \sum_{n=k}^{m} |F_{k,1;N,m}| \\ &= \quad |F_{k,1;N,m}| + O\left(\sum_{n=k+1}^{m} |F_{n,1;N,m}|\right) \\ &= \quad \frac{1}{k!}\frac{m^k}{N^k}\, e^{-m/N} + O\left(\sum_{n=k+1}^{m} \frac{1}{n!}\left(\frac{m}{N}\right)^n\right) \\ &= \quad \frac{1}{k!}\frac{m^k}{N^k}\, e^{-m/N} + O\left(\frac{m^{k+1}}{N^{k+1}}\right). \end{aligned} \tag{2.10}$$

Substituting into (2.7) yields

$$\begin{aligned} |E_{k;N,m}| \quad &\le \quad N \cdot |E_{k,1;N,m}| \\ &\le \quad \frac{1}{k!}\frac{m^k}{N^{k-1}}\, e^{-m/N} + O\left(\frac{m^{k+1}}{N^k}\right). \end{aligned} \tag{2.11}$$

If we take $m = N^{\frac{k-1}{k}}$ then the main term is of size $\frac{1}{k!}$ (as $e^{-m/N} = e^{-1/N^{1/k}} = 1 + O(N^{-1/k})$) and the error term is $O\left(N^{-1/k}\right)$. Thus we have shown for k fixed and $m = N^{\frac{k-1}{k}}$ that

$$|E_{k;N,m}| \ \le\ \frac{1}{k!} + O\left(N^{-1/k}\right), \tag{2.12}$$

completing the proof of the upper bound. □

Let $E_{n_1,i_1,n_2,i_2;N,m}$ be the event of at least n_1 balls in box i_1 and at least n_2 balls in box i_2 (with m balls in all, N boxes). By inclusion-exclusion we have

$$|E_{k;N,m}| \ >\ \sum_{i=1}^{N} |E_{k,i;N,m}| - \sum_{i_1=1}^{N-1} \sum_{i_2=i_1+1}^{N} |E_{k,i_1,k,i_2;N,m}|. \tag{2.13}$$

The left hand side, $|E_{k,i;N,m}|$, counts how many times at least one box has at least k balls. If this happens, then there must be at least one index i such that it is counted in an $E_{k,i;N,m}$. If there are two such indices, it is counted twice, but then we subtract it once from an $E_{k,i_1,k,i_2;N,m}$ term. If exactly $\ell \ge 2$ boxes contain at least k balls, then we have counted this ℓ times from the $E_{k,i;N,m}$ terms and subtracted it $\binom{\ell}{2}$ times from the $E_{k,i_1,k,i_2;N,m}$ terms. Thus (2.13) is a lower bound for $|E_{k;N,m}|$.

Proof of the Lower Bound in (1.5). Thus by the above arguments and symmetry, we need only compute a good estimate for $|E_{k,1,k,2;N,m}|$, as

$$|E_{k;N,m}| \ \ge\ N \cdot |E_{k,1;N,m}| - \frac{N(N-1)}{2} \cdot |E_{k,1,k,2;N,m}|. \tag{2.14}$$

Let $F_{n_1,1,n_2,2;N,m}$ be the event of exactly n_1 balls in the first box and exactly n_2 balls in the second box (with m balls and N boxes). Then for $m \gg \max(n_1, n_2)$,

$$|F_{n_1,1,n_2,2;N,m}| \ =\ \binom{m}{n_1}\frac{1}{N^{n_1}}\binom{m-n_1}{n_2}\frac{1}{N^{n_2}}\binom{m-n_1-n_2}{m-n_1-n_2}\left(1-\frac{1}{N}\right)^{m-n_1-n_2} + O(N^{-1}). \tag{2.15}$$

The main term is when $n_1 = n_2 = k$, which gives

$$|F_{k,1,k,2;N,m}| \ \sim\ \frac{1}{k!k!}\frac{m^{2k}}{N^{2k}}\ e^{-m/N}. \tag{2.16}$$

We bound the contribution from terms with each $n_i \ge k$ and $n_1 + n_2 \ge 2k+1$. If $n_1 + n_2 = \ell$, there are clearly only $\ell - 1$ pairs of positive integers (n_1, n_2) that sum to ℓ (of course, there are fewer pairs for us, as each must be at least k). As $\ell - 1 \le n_1 n_2$, we have

$$\sum_{\substack{n_1,n_2 \ge k\\ n_1+n_2 \ge 2k+1}} |F_{n_1,1,n_2,2;N,m}| \ =\ O\left(\sum_{\ell=2k+1}^{m} \frac{1}{\lfloor\frac{\ell-2}{2}\rfloor!}\left(\frac{m}{N}\right)^{\ell}\right) \ =\ O\left(\frac{m^{2k+1}}{N^{2k+1}}\right). \tag{2.17}$$

Therefore, we have

$$|E_{k,1,k,2;N,m}| \ =\ \frac{1}{k!k!}\frac{m^{2k}}{N^{2k}}\ e^{-m/N} + O\left(\frac{m^{2k+1}}{N^{2k+1}}\right). \tag{2.18}$$

Substituting into (2.14) and using (2.10) for the size of $|E_{k,1;N,m}|$ yields

$$\begin{aligned} |E_{k;N,m}| \quad &> \quad N \cdot \left[\frac{1}{k!}\frac{m^k}{N^k}\ e^{-m/N} + O\left(\frac{m^{k+1}}{N^{k+1}}\right)\right] \\ &\quad - \frac{N^2}{2} \cdot \left[\frac{1}{k!k!}\frac{m^{2k}}{N^{2k}}\ e^{-m/N} + O\left(\frac{m^{2k+1}}{N^{2k+1}}\right)\right]. \end{aligned} \tag{2.19}$$

Again taking $m = N^{\frac{k-1}{k}}$ (so $\frac{m^k}{N^{k-1}} = 1$), we find that

$$|E_{k;N,m}| \quad > \quad \frac{2\cdot k! - 1}{2\cdot k!^2} + O\left(N^{-1/k}\right), \tag{2.20}$$

completing the proof of the lower bound. □

Remark 2.1. Using the lower bound, we can bootstrap and ensure a high probability of having at least one box with at least k balls. The probability of *not* having at least k balls in one of the boxes is at most

$$1 - \frac{2\cdot k! - 1}{2\cdot k!^2} + O\left(N^{-1/k}\right), \tag{2.21}$$

remembering of course that $m = N^{\frac{k-1}{k}}$. Consider now a independent sets of $m = N^{\frac{k-1}{k}}$ balls. The probability that *none* of these a sets has at least one box with k balls is

$$\left(1 - \frac{2\cdot k! - 1}{2\cdot k!^2} + O\left(N^{-1/k}\right)\right)^a. \tag{2.22}$$

By choosing a sufficiently large, we can make this probability as close to zero as we like, or equivalently make the probability that if we take at least $aN^{\frac{k-1}{k}}$ balls then at least one box has at least k balls. By taking a to be a small power of N, we can make the probability 1 plus a smaller term.

Remark 2.2. Note in Remark 2.1 that we considered a independent sets of m balls. In finding our bounds of having at least k balls in a box we do not allow (say) $k - k'$ balls in box 1 from the first set and k' balls in box 1 from the second set; thus the a we take is almost surely much larger than needed.

3. LETTING k DEPEND ON N

We discuss what happens if we try to use these arguments with k growing with N. Specifically, if we have $m = N^{2-\epsilon}$, then is there a positive probability (as $N \to \infty$) of having at least one box with at least $k = N$ balls in it? We use Stirling's formula, which gives us the approximation

$$n! \ \sim \ n^n e^{-n}\sqrt{2\pi n}. \tag{3.23}$$

Let us first consider the probability of having at least N balls in the first box. The probability of exactly n balls in the first box is

$$\begin{aligned} |P_{n,1;N,m}| \quad &= \quad \binom{m}{n}\frac{1}{N^n}\binom{m-n}{m-n}\left(1-\frac{1}{N}\right)^{m-n} \\ &\leq \quad \frac{1}{n!}\frac{m^n}{N^n}\, e^{-(m-n)/N}. \end{aligned} \tag{3.24}$$

We first bound the contribution when $n \in \{N^{2-2\epsilon}, \dots, m\}$, where $m = N^{2-\epsilon}$. These contribute

$$\begin{aligned}
\sum_{n=N^{2-2\epsilon}}^{N^{2-\epsilon}} |P_{n,1;N,N^{2-\epsilon}}| &\le \sum_{n=N^{2-2\epsilon}}^{N^{2-\epsilon}} \frac{1}{n^n e^{-n}\sqrt{2\pi n}} \frac{m^n}{N^n}\ e^{-(m-n)/N} \\
&\ll \sum_{n=N^{2-2\epsilon}}^{N^{2-\epsilon}} (2\pi n)^{-\frac{1}{2}} \left(\frac{em}{nN}\right)^n e^{-(m-n)/N} \\
&\ll \sum_{n=N^{2-2\epsilon}}^{N^{2-\epsilon}} (2\pi n)^{-\frac{1}{2}} \left(\frac{eN^{2-\epsilon}}{nN}\right)^n \\
&\ll \sum_{n=N^{2-2\epsilon}}^{N^{2-\epsilon}} n^{-\frac{1}{2}} \left(\frac{e}{N^{1-\epsilon}}\right)^{N^{2-2\epsilon}} \\
&\ll \sum_{n=N^{2-2\epsilon}}^{N^{2-\epsilon}} n^{-\frac{1}{2}} e^{N^{2-2\epsilon}\log(e/N^{1-\epsilon})} \\
&\ll N^{1-\frac{\epsilon}{2}} e^{-(1-\epsilon)N^{2-2\epsilon}\log N + N^{2-2\epsilon}} \\
&\ll e^{-(1-\epsilon)N^{2-2\epsilon}\log N + N^{2-2\epsilon} + (1-\frac{\epsilon}{2})\log N}. \qquad (3.25)
\end{aligned}$$

We consider the contribution from terms with $n \in \{N, \dots, N^{2-2\epsilon}\}$; note $n \le m = N^{2-\epsilon}$. For such n we have (δ a positive constant below) that

$$\begin{aligned}
\sum_{n=N}^{N^{2-2\epsilon}} |P_{n,1;N,N^{2-\epsilon}}| &\le \sum_{n=N}^{N^{2-2\epsilon}} \frac{1}{n^n e^{-n}\sqrt{2\pi n}} \frac{m^n}{N^n}\ e^{-(m-n)/N} \\
&\ll \sum_{n=N}^{N^{2-2\epsilon}} (2\pi n)^{-\frac{1}{2}} \left(\frac{em}{nN}\right)^n e^{-(m-n)/N} \\
&\ll \sum_{n=N}^{N^{2-2\epsilon}} (2\pi n)^{-\frac{1}{2}} \left(\frac{eN^{2-\epsilon}}{nN}\right)^n e^{-(N^{2-\epsilon}-n)/N} \\
&\ll \sum_{n=N}^{N^{2-2\epsilon}} n^{-\frac{1}{2}} \left(\frac{e}{N^{\epsilon}(n/N)}\right)^n e^{-\delta N^{1-\epsilon}} \\
&\ll \sum_{n=N}^{N^{2-2\epsilon}} n^{-\frac{1}{2}} \left(\frac{e}{N^{\epsilon}}\right)^N e^{-\delta N^{1-\epsilon}} \\
&\ll \sum_{n=N}^{N^{2-2\epsilon}} n^{-\frac{1}{2}} e^{N\log(e/N^{\epsilon})} e^{-\delta N^{1-\epsilon}} \\
&\ll \sum_{n=N}^{N^{2-2\epsilon}} n^{-\frac{1}{2}} e^{-\epsilon N\log N + N - \delta N^{1-\epsilon}} \\
&\ll N^{1-\epsilon} e^{-\epsilon N\log N + N - \delta N^{1-\epsilon}} \\
&\ll e^{-\epsilon N\log N + N - \delta N^{1-\epsilon} + (1-\epsilon)\log N}. \qquad (3.26)
\end{aligned}$$

Thus from (3.25) and (3.26) we have

$$\sum_{n=N}^{N^{2-\epsilon}} |P_{n,1;N,N^{2-\epsilon}}| \ll e^{-\epsilon N \log N + N - \delta N^{1-\epsilon} + (1-\epsilon)\log N} + e^{-(1-\epsilon)N^{2-2\epsilon}\log N + N^{2-2\epsilon} + (1-\frac{\epsilon}{2})\log N}. \tag{3.27}$$

For $m = N^{2-\epsilon}$, as

$$E_{N;N,N^{2-\epsilon}} \subset \bigcup_{i=1}^{N} \bigcup_{n=N}^{N^{2-\epsilon}} P_{n,i;N,N^{2-\epsilon}}, \tag{3.28}$$

we finally obtain that

$$\begin{aligned} |E_{N;N,N^{2-\epsilon}}| &\ll N \cdot e^{-\epsilon N \log N + N - \delta N^{1-\epsilon} + (1-\epsilon)\log N} \\ &\quad + N \cdot e^{-(1-\epsilon)N^{2-2\epsilon}\log N + N^{2-2\epsilon} + (1-\frac{\epsilon}{2})\log N} \\ &\ll e^{-\epsilon N \log N + N - \delta N^{1-\epsilon} + (2-\epsilon)\log N} \\ &\quad + e^{-(1-\epsilon)N^{2-2\epsilon}\log N + N^{2-2\epsilon} + (2-\frac{\epsilon}{2})\log N}, \end{aligned} \tag{3.29}$$

which yields

Theorem 3.1. *There is negligible probability of having at least N balls in one of N boxes when there are $N^{2-\epsilon}$ balls.*

Remark 3.2. As $\delta \in (0, 1]$, even if we were to take

$$\epsilon = \frac{\theta}{\log N} \tag{3.30}$$

in the above arguments (for some $\theta > 1$), we would still have $|E_{N;N,N^{2-\epsilon}}| = o(1)$ for such m.

4. Moment Arguments

Let's analyze the mean and standard deviations when m independent balls are tossed into N boxes (each box equally likely). Let $w_{i,1}$ be the binary indicator variable for ball i and box 1. Thus $w_{i,1}$ is 1 with probability $p = \frac{1}{N}$ and 0 with probability $q = 1 - \frac{1}{N}$. Note the mean of $w_{i,1}$ is $\frac{1}{N}$ and the standard deviation is $\sqrt{pq}$, which is approximately $N^{-\frac{1}{2}}$.

If we let $w_1 = \sum_{i=1}^{m} w_{i,1}$, then the mean is simply $\frac{m}{N}$ and the standard deviation is $\sqrt{mpq}$.

If we fix k, we've seen we need to take $m \sim N^{\frac{k-1}{k}}$. Such a choice leads to the expected number of balls in the first box of $\frac{m}{N} = N^{-1/k}$, with a standard deviation of $\sqrt{mpq} \sim N^{-1/2k}$. Thus we need to be on the order of $kN^{1/2k}$ standard deviations from the mean; of course, we have N boxes and need this just for *one* box. We can look at this in terms of m – we need on the order of $k\, m^{1/2(k-1)}$ standard deviations.

If we let $k = N$ and $m = N^{2-\epsilon}$, then the expected number of balls in the first box is $\frac{m}{N} = N^{1-\epsilon}$, and the standard deviation is $\sqrt{mpq} \sim N^{\frac{1}{2}-\frac{\epsilon}{2}}$. Thus we would need on the order of $N^{\frac{1}{2}-\frac{\epsilon}{2}}$ standard deviations from the mean (we need to get up to N, each standard deviation adds about $N^{\frac{1}{2}-\frac{\epsilon}{2}}$ so we need $N^{\frac{1}{2}+\frac{\epsilon}{2}}$ such steps); of course, we have N boxes and this is just for *one* box. We can look at this in terms of m – we need on the order of $m^{\frac{1}{4}+\epsilon'}$ standard deviations.

The above arguments are meant to try and provide some insight as to what breaks down when we consider $k = N$ and $m = N^{2-\epsilon}$. These are just some quick thoughts.

DEPARTMENT OF MATHEMATICS, BROWN UNIVERSITY, 151 THAYER STREET, PROVIDENCE, RI 02912
E-mail address: sjmiller@math.brown.edu