Bernoulli numbers and the probability of a birthday surprise ⋆

Boaz Tsaban

Department of Mathematics and Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel

Abstract

A birthday surprise is the event that, given k uniformly random samples from a sample space of size n, at least two of them are identical. We show that Bernoulli numbers can be used to derive arbitrarily exact bounds on the probability of a birthday surprise. This result can be used in arbitrary precision calculators, and it can be applied to better understand some questions in communication security and pseudorandom number generation.

Key words: birthday paradox, power sums, Bernoulli numbers, arbitrary precision calculators, pseudorandomness

1 Introduction

In this note we address the probability β_n^k that in a sample of k uniformly random elements out of a space of size n there exist at least two identical elements. This problem has a long history and a wide range of applications. The term *birthday surprise* for a collision of (at least) two elements in the sample comes from the case $n = 365$, where the problem can be stated as follows: Assuming that the birthday of people distributes uniformly over the year, what is the probability that in a class of k students, at least two have the same birthday?

⋆ Dedicated to my wife Lea on her birthday

Email address: tsaban@macs.biu.ac.il (Boaz Tsaban).
URL: http://www.cs.biu.ac.il/~tsaban (Boaz Tsaban).

It is clear (and well known) that the expected number of collisions (or birthdays) in a sample of k out of n is:

$$\binom{k}{2}\frac{1}{n} = \frac{k(k-1)}{2n}.$$

(Indeed, for each distinct i and j in the range $\{1,\ldots,k\}$, let X_{ij} be the random variable taking the value 1 if samples i and j obtained the same value and 0 otherwise. Then the expected number of collisions is $E(\sum_{i\neq j} X_{ij}) = \sum_{i\neq j} E(X_{ij}) = \sum_{i\neq j} \frac{1}{n} = \binom{k}{2}\frac{1}{n}$.)

Thus, 28 students are enough to make the expected number of common birthdays greater than 1. This seemingly surprising phenomenon has got the name *birthday surprise*, or *birthday paradox*.

In several applications, it is desirable to have exact bounds on the probability of a collision. For example, if some electronic application chooses pseudorandom numbers as passwords for its users, it may be a *bad* surprise if two users get the same password by coincidence. It is this term "by coincidence" that we wish to make precise.

2 Bounding the probability of a birthday surprise

When k and n are relatively small, it is a manner of simple calculation to determine β_n^k. The probability that all samples are distinct is:

$$\pi_n^k = \left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right), \tag{1}$$

and $\beta_n^k = 1-\pi_n^k$. For example, one can check directly that $\beta_{365}^{23} > 1/2$, that is, in a class of 23 students the probability that two share the same birthday is greater than 1/2. This is another variant of the *Birthday surprise.*[1]

The calculation becomes problematic when k and n are large, both due to precision problems and computational complexity (in cryptographic applications k may be of the order of trillions, i.e., thousands of billions). This problem can be overcome by considering the logarithm of the product:

$$\ln(\pi_n^k) = \sum_{i=1}^{k-1} \ln\left(1-\frac{i}{n}\right).$$

[1] To experience this phenomenon experimentally, the reader is referred to [8].

Since each i is smaller than n, we can use the Taylor expansion $\ln(1-x) = -\sum_{m=1}^{\infty} x^m/m$ ($|x|<1$) to get that

$$-\ln(\pi_n^k) = \sum_{i=1}^{k-1}\sum_{m=1}^{\infty}\frac{(i/n)^m}{m} = \sum_{m=1}^{\infty}\frac{1}{mn^m}\sum_{i=1}^{k-1} i^m. \qquad (2)$$

(Changing the order of summation is possible because the sums involve positive coefficients.)

The coefficients $\mathfrak{p}(k-1,m) := \sum_{i=1}^{k-1} i^m$ (which are often called *sums of powers*, or simply *power sums*) play a key role in our estimation of the birthday probability. Efficient calculations of the first few power sums go back to ancient mathematics. [2] In particular, we have: $\mathfrak{p}(k,1) = k(k+1)/2$, and $\mathfrak{p}(k,2) = k(k+1)(2k+1)/6$. Higher order power sums can be found recursively using *Bernoulli numbers.*

The Bernoulli numbers (which are indexed by superscripts) $1 = B^0$, B^1, B^2, B^3, B^4,... are defined by the formal equation "$B^n = (B-1)^n$" for $n > 1$, where the quotation marks indicate that the involved terms are to be expanded in formal powers of B before interpreting. Thus:

- $B^2 = B^2 - 2B^1 + 1$, whence $B^1 = 1/2$,
- $B^3 = B^3 - 3B^2 + 3B^1 - 1$, whence $B^2 = 1/6$,

etc. We thus get that $B^3 = 0$, $B^4 = -1/30$, $B^5 = 0$, $B^6 = 1/42$, $B^7 = 0$, and so on. It follows that for each m,

$$\mathfrak{p}(k,m) = \frac{\text{“}(k+B)^{m+1} - B^{m+1}\text{”}}{m+1}$$

(*Faulhaber's formula* [6].) Thus, the coefficients $\mathfrak{p}(k,m)$ can be efficiently calculated for small values of m. In particular, we get that

- $\mathfrak{p}(k,3) = \frac{1}{4}k^4 + \frac{1}{2}k^3 + \frac{1}{4}k^2$,
- $\mathfrak{p}(k,4) = \frac{1}{5}k^5 + \frac{1}{2}k^4 + \frac{1}{3}k^3 - \frac{1}{30}k$,
- $\mathfrak{p}(k,5) = \frac{1}{6}k^6 + \frac{1}{2}k^5 + \frac{5}{12}k^4 - \frac{1}{12}k^2$,
- $\mathfrak{p}(k,6) = \frac{1}{7}k^7 + \frac{1}{2}k^6 + \frac{1}{2}k^5 - \frac{1}{6}k^3 + \frac{1}{42}k$,
- $\mathfrak{p}(k,7) = \frac{1}{8}k^8 + \frac{1}{2}k^7 + \frac{7}{12}k^6 - \frac{7}{24}k^4 + \frac{1}{12}k^2$,

etc. In order to show that this is enough, we need to bound the tail of the series in Equation 2. We will achieve this by effectively bounding the power sums.

[2] Archimedes (ca. 287-212 BCE) provided a geometrical derivation of a "formula" for the sum of squares [9].

Lemma 1 *Let k be any natural number, and assume that $f : (0,k) \to \mathbb{R}^+$ is such that $f''(x)$ exists, and is nonnegative for all $x \in (0,k)$. Then:*

$$\sum_{i=1}^{k} f(i) < \int_{0}^{k} f(x+\tfrac{1}{2})dx.$$

Proof. For each interval $[i, i+1]$ $(i = 0, \ldots, k-1)$, the tangent to the graph of $f(x+\frac{1}{2})$ at $x = i+\frac{1}{2}$ goes below the graph of $f(x+\frac{1}{2})$. This implies that the area of the added part is greater than that of the uncovered part. $\square$

Using Lemma 1, we have that for all $m > 1$,

$$\sum_{i=1}^{k-1} i^m < \int_{0}^{k-1} (x+\tfrac{1}{2})^m dx < \frac{(k-\frac{1}{2})^{m+1}}{m+1}.$$

Thus,

$$\sum_{m=N}^{\infty} \frac{\mathfrak{p}(k-1,m)}{mn^m} < \sum_{m=N}^{\infty} \frac{(k-\frac{1}{2})^{m+1}}{m(m+1)n^m} < \frac{k-\frac{1}{2}}{N(N+1)} \sum_{m=N}^{\infty} \left(\frac{k-\frac{1}{2}}{n}\right)^m =$$

$$= \frac{k-\frac{1}{2}}{N(N+1)} \cdot \frac{\left(\frac{k-\frac{1}{2}}{n}\right)^N}{1-\frac{k-\frac{1}{2}}{n}} = \frac{(k-\frac{1}{2})^{N+1}}{N(N+1)\left(1-\frac{k-\frac{1}{2}}{n}\right)n^N}. \quad (3)$$

We thus have the following.

Theorem 2 *Let π_n^k denote the probability that all elements in a sample of k elements out of n are distinct. For a natural number N, define*

$$\epsilon_n^k(N) := \frac{(k-\frac{1}{2})^{N+1}}{N(N+1)\left(1-\frac{k-\frac{1}{2}}{n}\right)n^N}.$$

Then

$$\sum_{m=1}^{N-1} \frac{\mathfrak{p}(k-1,m)}{mn^m} < -\ln(\pi_n^k) < \sum_{m=1}^{N-1} \frac{\mathfrak{p}(k-1,m)}{mn^m} + \epsilon_n^k(N).$$

For example, for $N = 2$ we get:

$$\frac{(k-1)k}{2n} < -\ln(\pi_n^k) < \frac{(k-1)k}{2n} + \frac{(k-\frac{1}{2})^3}{6n^2\left(1-\frac{k-\frac{1}{2}}{n}\right)}.$$

We demonstrate the tightness of these bounds with a few concrete examples:

Example 3 Let us bound the probability that in a class of 5 students there exist two sharing the same birthday. Using Theorem 2 with $N = 2$ we get by simple calculation that $\frac{2}{73} < -\ln(\pi^5_{365}) < \frac{2}{73} + \frac{243}{2105320}$, or numerically,[3] $0.0273972 < -\ln(\pi^5_{365}) < 0.0275127$. Thus, $0.0270253 < \beta^5_{365} < 0.0271377$. Repeating the calculations with $N = 3$ yields $0.0271349 < \beta^5_{365} < 0.0271356$. $N = 4$ shows that $\beta^5_{365} = 0.0271355\ldots$.

Example 4 We bound the probability that in a class of 73 students there exist two sharing the same birthday, using $N = 2$: $\frac{36}{5} < -\ln(\pi^{73}_{365}) < \frac{36}{5} + \frac{121945}{255792}$, and numerically we get that $0.9992534 < \beta^{73}_{365} < 0.9995882$. For $N = 3$ we get $0.9995365 < \beta^{73}_{365} < 0.9995631$, and for $N = 8$ we get that $\beta^{73}_{365} = 0.9995608\ldots$.

In Theorem 2, $\epsilon^k_n(N)$ converges to 0 exponentially fast with N. In fact, the upper bound is a very good approximation to the actual probability, as can be seen in the above examples. The reason for this is the effectiveness of the bound in Lemma 1 (see [4] for an analysis of this bound as an approximation).

For $k < \sqrt{n}$, we can bound β^k_n directly: Note that for $|x| < 1$ and odd M, $\sum_{m=0}^{M}(-x)^m/m! < e^{-x} < \sum_{m=0}^{M+1}(-x)^m/m!$.

Corollary 5 *Let β^k_n denote the probability of a birthday surprise in a sample of k out of n, and let $l_N(k,n)$ and $u_N(k,n)$ be the lower and upper bounds from Theorem 2, respectively. Then for all odd M,*

$$-\sum_{m=1}^{M+1}\frac{(-l_N(k,n))^m}{m!} < \beta^k_n < -\sum_{m=1}^{M}\frac{(-u_N(k,n))^m}{m!}.$$

For example, when $M = 1$ we get that

$$\frac{(k-1)k}{2n} - \frac{(k-1)^2k^2}{4n^2} < \beta^k_n < \frac{(k-1)k}{2n} + \frac{(k-\frac{1}{2})^3}{6n^2\left(1-\frac{k-\frac{1}{2}}{n}\right)}. \tag{4}$$

The explicit bounds become more complicated when $M > 1$, but once the lower and upper bounds in Theorem 2 are computed numerically, bounding β^k_n using Corollary 5 is easy. However, Corollary 5 is not really needed in order to deduce the bounds – these can be calculated directly from the bounds of Theorem 2, e.g. using the exponential function built in calculators.

[3] All calculations in this paper were performed using the GNU bc calculator [5], with a scale of 500 digits.

Remark 6 (1) It can be proved directly that in fact $\beta_n^k < \frac{(k-1)k}{2n}$ [2]. However, it is not clear how to extend the direct argument to get tighter bounds in a straightforward manner.

(2) Our lower bound in Equation 4 compares favorably with the lower bound $\left(1-\frac{1}{e}\right)\frac{(k-1)k}{2n}$ from [2] when $k \leq \sqrt{2n/e}$ (when $k > \sqrt{2n/e}$ we need to take larger values of M to get a better approximation).

(3) $\mathfrak{p}(k-1, m)$ is bounded from below by $(k-1)^{m+1}/(m+1)$. This implies a slight improvement on Theorem 2.

3 Some applications

3.1 Arbitrary percision calculators

Arbitrary percision calculators do calculations to any desired level of accuracy. Well-known examples are the *bc* and *GNU bc* [5] calculators. Theorem 2 allows calculting β_n^k to any desired level of accuracy (in this case, the parameter N will be determined by the required level of accuracy), and in practical time. An example of such calculation appears below (Example 7).

3.2 Cryptography

The probability of a birthday surprise plays an important role in the security analysis of various cryptographic systems. For this purpose, it is common to use the approximation $\beta_n^k \approx k^2/2n$. However, in *concrete security* analysis it is prefered to have exact bounds rather than estimations (see [1] and references therein).

The second item of Remark 6 implies that security bounds derived using earlier methods are tighter than previously thought. The following example demonstrates the tightness of the bounds of Theorem 2 for these purposes.

Example 7 In [3], $\beta_{2^{128}}^{2^{32}}$ is estimated approximately. Using Theorem 2 with $N = 2$, we get that in fact,

$$2^{-65.00000000033590361502507960391 03} < \beta_{2^{128}}^{2^{32}} < 2^{-65.0000000003359036150250796039042}.$$

With $N = 3$ we get that $\beta_{2^{128}}^{2^{32}}$ lies between

$$2^{-65.000000000335903615025079603904203942942489665995829764250752}$$

and

$$2^{-65.0000000000335903615025079603904203942942489665995829764250713}.$$

The remarkable tightness of these bounds is due to the fact that 2^{32} is much smaller than 2^{128}.

Another application of our results is for estimations of the quality of approximations such as $\binom{n}{k} \approx n^k/k!$ (when $k << n$):

Fact 8 $\binom{n}{k} = \frac{n^k}{k!} \cdot \pi_n^k$.

Thus the quality of this approximation is directly related to the quality of the approximation $\pi_n^k \approx 1$, which is well understood via Theorem 2.

π_n^k appears in many other natural contexts. For example, assume that a function $f : \{0, \ldots, n-1\} \to \{0, \ldots, n-1\}$ is chosen with uniform probability from the set of all such functions, and fix an element $x \in \{0, \ldots, n-1\}$. Then we have the following immediate observation.

Fact 9 *The probability that the orbit of x under f has size exactly k is $\pi_n^k \cdot \frac{k}{n}$. The probability that the size of the orbit of x is larger than k is simply π_n^k.*

These probabilities play an important role in the theory of iterative pseudorandom number generation (see [10] for a typical example).

4 Final remarks and acknowledgments

For a nice account of power sums see [7]. An accessible presentation and proof of Faulhaber's formula appears in [6]. The author thanks John H. Conway for the nice introduction to Bernoulli numbers, and Ron Adin for reading this note and detecting some typos.

References

[1] Mihir Bellare, *Practice-oriented provable-security*, Lecture Notes in Computer Science **1561** (1999), pp. 1–15.

[2] Mihir Bellare, Joe Kilian, and Phillip Rogaway, *The security of the Cipher Block Chaining message authentication code*, Journal of Computer and System Sciences **61** (2000), pp. 362–399.

[3] Mihir Bellare, Oded Goldreich, and Hugo Krawczyk, *Stateless evaluation of pseudorandom functions: Security beyond the birthday barrier*, Advances in Cryptology – CRYPTO 99 Proceedings (ed. M. Wiener), Lecture Notes in Computer Science **1666** (1999), pp. 270–287.

[4] Brian L. Burrows and Richard F. Talbot, *Sums of powers of integers*, American Mathematical Monthly **91** (1986), 394–403.

[5] GNU BC, `ftp://prep.ai.mit.edu/pub/gnu/bc/`.

[6] John H. Conway and Richard K. Guy, *The Book of Numbers*, Copernicus (Springer-Verlag), New York: 1996, 106–109.

[7] A. W. F. Edwards, *A quick route to sums of powers*, American Mathematical Monthly **93** (1986), 451–455.

[8] `http://www-stat.stanford.edu/~susan/surprise/`

[9] Victor J. Katz, *A History of Mathematics: An Introduction*, New York: HarperCollins College Publishers, 1993, 106.

[10] Adi Shamir and Boaz Tsaban, *Guaranteeing the diversity of pseudorandom generators*, Information and Computation **171** (2001), 350–363. `http://arxiv.org/abs/cs.CR/0112014`