Majorization and the Birthday Inequality

M. LAWRENCE CLEVENSON
WILLIAM WATKINS
California State University
Northridge, CA 91330

Introduction If there are k people in a room, what is the probability that at least two of them have the same birthday? Intuitively, people guess that the answer is $k/365$ or close to it [4]. Of course the probability of a match is

$$1 - (365)(364) \cdots (365 - k + 1)/365^k,$$

which is greater than 0.5 even if there are only 23 people in the room. Probability instructors are fond of demonstrating this fact in class by asking the students to announce their birthdays until a match is discovered. The authors are courageous enough to try the demonstration in classes with 30 or more students so that the probability of a match is at least 0.706.

Recently, however, we got worried. The probabilities given above are based on the assumption that a randomly chosen individual is just as likely to have a birthday on, say, October 1 as on January 10. In other words, the assumption is that

$$p_i = 1/365,$$

for $i = 1, \ldots, 365$, where p_i is the probability that a randomly chosen individual has a birthday on the ith day of the year. Since some births are scheduled, this assumption of equally likely birthdays is questionable. So, the probability of a match with, say, 30 people in the room is not necessarily 0.706. Could it be smaller? Should we be less courageous in our classroom demonstrations? The answer is no. If the distribution of birthdays is not uniform, then the probability of a match is at least as large as it is when the birthdays are uniformly distributed throughout the year. We call this fact the birthday inequality.

It isn't new. Bloom [1] and Rust [6] give proofs using Lagrange multipliers and Munford [5] gives one using elementary symmetric functions. Our purpose is to strengthen the birthday inequality and to place it in its natural setting—the theory of majorization and Schur-convexity. The stronger result appears in Marshall and Olkin's book on majorization [3, p. 305] as a corollary to a much more general theorem. Our treatment is elementary.

The birthday inequality Let the n-tuple $\mathbf{p} = (p_1, \ldots, p_n)$ be any probability vector of birthdays for the $n = 365$ days of the year. Given $\mathbf{p}$, let $P_k(\mathbf{p})$ denote the probability of at least one match among k people and let $Q_k(\mathbf{p}) = 1 - P_k(\mathbf{p})$ be the probability of the complementary event that all k birthdays are different. We want to prove $P_k(\mathbf{p}) \geqslant P_k(\mathbf{u})$, or equivalently

$$Q_k(\mathbf{p}) \leqslant Q_k(\mathbf{u}), \tag{1}$$

where $\mathbf{u} = (1/n, \ldots, 1/n)$ is the uniform distribution of birthdays.

We need a formula for $Q_k(\mathbf{p})$. Given k people and an ordered selection $i_1, \ldots, i_k$ of k days of the year, the probability that person j is born on day i_j, for $j = 1, \ldots, k$ is $p_{i_1} \cdots p_{i_k}$. It follows that

$$Q_k(\mathbf{p}) = \sum p_{i_1} \cdots p_{i_k}, \tag{2}$$

where the sum is taken over all sequences $i_1, \ldots, i_k$ of distinct integers from 1 to n. This sum is easy to compute for the uniform distribution. There are $(365)(364) \cdots (365 - k + 1)$ ways to pick a sequence of k distinct integers from $1, \ldots, 365$. So the probability of no match is given by $Q_k(\mathbf{u}) = (365)(364) \cdots (365 - k + 1)/365^k$, as we saw earlier.

The proof of inequality (1) requires a slight reformulation of formula (2) in terms of elementary symmetric functions. The *kth elementary symmetric function* of $\mathbf{p} = (p_1, \ldots, p_n)$ is defined by

$$E_k(\mathbf{p}) = \sum p_{i_1} \cdots p_{i_k},$$

where the sum is taken over all strictly increasing sequences $i_1, \ldots, i_k$ of integers from 1 to n. Since $Q_k(\mathbf{p}) = k! E_k(\mathbf{p})$, it suffices to prove that

$$E_k(\mathbf{p}) \leqslant E_k(\mathbf{u}) \tag{3}$$

instead of the original birthday inequality (1).

Proof of the birthday inequality One of the ideas involved in the proof is that of a *T-transform*. Given an n-tuple, say $\mathbf{p} = (.28, .16, .41, .11, .04)$, pick two coordinates, say $p_1 = .28$ and $p_2 = .16$. A T-transform has the following action on $\mathbf{p}$:

$$(.28, .16, .41, .11, .04) \to (.28 - d, .16 + d, .41, .11, .04),$$

where d is a positive number satisfying $d \leqslant |p_1 - p_2|$ that must be subtracted from the larger of the two coordinates and added to the smaller. If we choose $d = .08$, then the corresponding T-transform sends $\mathbf{p}$ to $(.20, .24, .41, .11, .04)$. Only two of the coordinates of $\mathbf{p}$ are changed and they get closer together. We might say that $(.20, .24, .41, .11, .04)$ is more uniform or less spread out than $(.28, .16, .41, .11, .04)$—an idea that we will develop in the next section. The T-transform does not change the sum of the coordinates of $\mathbf{p}$.

An appropriate sequence of T-transforms can change any probability vector $\mathbf{p}$ into the uniform probability vector $\mathbf{u}$. To see this observe that if $\mathbf{p} \neq \mathbf{u}$ then $\mathbf{p}$ has a pair of coordinates p_i, p_j, satisfying $p_i > 1/n > p_j$. Let T be the T-transform that subtracts $d = p_i - 1/n$ from the ith coordinate of $\mathbf{p}$ and adds d to the jth coordinate. Then the ith coordinate of $T\mathbf{p}$ is $1/n$. In this way $\mathbf{p}$ can be transformed by a sequence of T-transforms to $\mathbf{u}$. In our example, four T-transforms are needed to get from $\mathbf{p}$ to $\mathbf{u}$:

$$\begin{aligned}(.28, .16, .41, .11, .04) &\to (.20, .24, .41, .11, .04)\\ &\to (.20, .20, .41, .15, .04)\\ &\to (.20, .20, .20, .15, .25)\\ &\to (.20, .20, .20, .20, .20).\end{aligned}$$

Why introduce the T-transform? Because it increases the value of the kth elementary symmetric function. That is, if $\mathbf{p}$ is a probability vector and T is a T-transform, then

$$E_k(\mathbf{p}) \leqslant E_k(T\mathbf{p}). \tag{4}$$

This is just what we need to prove the birthday inequality. If $\mathbf{p} \to \mathbf{q} \to \mathbf{r} \to \cdots \to \mathbf{u}$ is a sequence of T-transforms that takes $\mathbf{p}$ to the uniform distribution $\mathbf{u}$, then from (4) we have

$$E_k(\mathbf{p}) \leqslant E_k(\mathbf{q}) \leqslant E_k(\mathbf{r}) \leqslant \cdots \leqslant E_k(\mathbf{u})$$

and inequality (3) will be proved. All that remains in the proof of the birthday inequality is to show (4). We accomplish this using two properties of the elementary symmetric functions.

First, elementary symmetric functions are symmetric. That is, a permutation of the coordinates of $\mathbf{p}$ does not change the value of the elementary symmetric function. Thus to prove inequality (4), we may assume that $\mathbf{p} = (p_1, \ldots, p_n)$ with $p_1 > p_2$ and that the T-transform changes only the first two coordinates of $\mathbf{p}$. So suppose $T\mathbf{p} = (p_1 - d, p_2 + d, p_3, \ldots, p_n)$. The second property of the elementary symmetric functions is that they occur as the coefficients of the polynomial $(x + p_1)(x + p_2) \cdots (x + p_n)$. That is,

$$(x + p_1) \cdots (x + p_n) = x^n + E_1(\mathbf{p})x^{n-1} + \cdots + E_k(\mathbf{p})x^{n-k} + \cdots + E_n(\mathbf{p}).$$

Only the first two factors of the polynomials

$$(x + p_1)(x + p_2)(x + p_3) \cdots (x + p_n) \tag{5}$$

and

$$(x + p_1 - d)(x + p_2 + d)(x + p_3) \cdots (x + p_n) \tag{6}$$

differ. Examine the coefficients of the quadratics

$$(x + p_1)(x + p_2) = x^2 + (p_1 + p_2)x + p_1p_2 \tag{7}$$

and

$$(x + p_1 - d)(x + p_2 + d) = x^2 + (p_1 + p_2)x + \big(p_1p_2 + (p_1 - p_2)d - d^2\big). \tag{8}$$

Both coefficients of x^2 equal 1. The coefficients of x are also the same but the constant term in (7) is less than or equal to the constant term in (8) because $0 < d \leqslant p_1 - p_2$. Since the p_i are nonnegative, every coefficient of polynomial (5) is less than or equal to the corresponding coefficient in (6). Specifically, $E_k(\mathbf{p}) \leqslant E_k(T\mathbf{p})$. This completes the proof of (4) and the birthday inequality.

We have proved more than the birthday inequality. Namely, if $\mathbf{p}$ and $\mathbf{q}$ are probability vectors and there is a sequence of T-transforms changing $\mathbf{p}$ to $\mathbf{q}$, then $P_k(\mathbf{q}) \leqslant P_k(\mathbf{p})$. But for a given pair of probability vectors, it may not be easy to tell if one of them can be obtained by applying a sequence of T-transforms to the other one. In the next section we examine the conditions under which an arbitrary n-tuple $\mathbf{y}$ can be obtained from an n-tuple $\mathbf{x}$ by a sequence of T-transforms.

Majorization and Schur-convexity Let $\mathbf{x}$ and $\mathbf{y}$ be any two real n-tuples. How can we tell if there is a sequence of T-transforms that change $\mathbf{x}$ to $\mathbf{y}$? The following theorem [3, p. 7] provides a simple answer in terms of partial sums of the largest coordinates of the vectors $\mathbf{x}$ and $\mathbf{y}$. We omit the proof. The coordinates for any $\mathbf{z} = (z_1, \ldots, z_n)$, arranged in nonincreasing order will be denoted by $z_{[1]} \geqslant z_{[2]} \geqslant \cdots \geqslant z_{[n]}$.

THEOREM. *Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *and* $\mathbf{y} = (y_1, \ldots, y_n)$ *be real n-tuples. There exists a sequence of T-transforms that changes* $\mathbf{x}$ *to* $\mathbf{y}$ *if and only if*

$$x_{[1]} + \cdots + x_{[r]} \geqslant y_{[1]} + \cdots + y_{[r]}, \tag{9}$$

for all $r = 1, \ldots, n$, *and*

$$x_1 + \cdots + x_n = y_1 + \cdots + y_n. \tag{10}$$

These conditions are very easy to check for any given pair of n-tuples. For example, (9) and (10) hold for the vectors $\mathbf{x} = (.28, .16, .41, .11, .04)$ and $\mathbf{y} = (.20, .20, .20, .15, .25)$. Thus the theorem tells us what we have already verified—there is a sequence of T-transforms taking $\mathbf{x}$ to $\mathbf{y}$. If (9) and (10) hold, we say that $\mathbf{x}$ *majorizes* $\mathbf{y}$ and we write $\mathbf{x} \gg \mathbf{y}$. Intuitively, $\mathbf{x}$ is more spread out than $\mathbf{y}$ or equivalently, $\mathbf{y}$ is more uniform than $\mathbf{x}$. Equation (10) means that we can compare (via majorization) only those n-tuples whose coordinate sums are equal. The largest and smallest n-tuples with nonnegative coordinates and given coordinate sum S are $(S, 0, \ldots, 0)$ and $(S/n, \ldots, S/n)$, respectively. That is, if each $x_i \geqslant 0$ and $x_1 + \cdots + x_n = S$, then

$$(S, 0, \ldots, 0) \gg (x_1, \ldots, x_n) \gg (S/n, \ldots, S/n).$$

In the case of probability vectors, the uniform distribution $(1/n, \ldots, 1/n)$ is majorized by all other probability vectors.

We turn now to the definition of Schur-convex and Schur-concave functions. A real-valued function $F(\mathbf{x})$ of n real variables $\mathbf{x} = (x_1, \ldots, x_n)$ is *Schur-convex* if $\mathbf{x} \gg \mathbf{y}$ implies $F(\mathbf{x}) \geqslant F(\mathbf{y})$ and *Schur-concave* if $\mathbf{x} \gg \mathbf{y}$ implies $F(\mathbf{x}) \leqslant F(\mathbf{y})$. Equivalently, $F(\mathbf{x})$ is Schur-convex (Schur-concave) if $F(\mathbf{x}) \geqslant F(T\mathbf{x})$ ($F(\mathbf{x}) \leqslant F(T\mathbf{x})$), for all T-transforms. Perhaps Schur-increasing and Schur-decreasing would be better terms, but the names convex and concave have been in use for a long time. From the definition we see that among all n-tuples with nonnegative coordinates and a given coordinate sum S, a Schur-convex function takes its maximum value at $(S, 0, \ldots, 0)$ and its minimum value at $(S/n, \ldots, S/n)$. In the case of a probability vector, a Schur-convex function takes its maximum value at $(1, 0, \ldots, 0)$ and its minimum value at the uniform probability vector $\mathbf{u} = (1/n, \ldots, 1/n)$. We have shown that $P_k(\mathbf{p})$ is a Schur-convex function of $\mathbf{p}$. Thus the probability of matching birthdays is least when the probability vector is $\mathbf{u}$ and greatest when everyone is born on the same day of the year. For a Schur-concave function, the maximum and minimum are reversed.

Many standard inequalities involve Schur-convex or Schur-concave functions F and have the form

$$F(x_1, \ldots, x_n) \geqslant F(S/n, \ldots, S/n)$$

or

$$F(x_1, \ldots, x_n) \leqslant F(S/n, \ldots, S/n).$$

The arithmetic-geometric mean inequality is an example. For nonnegative numbers $x_1, \ldots, x_n$, the arithmetic-geometric mean inequality can be written in the form

$$x_1 \ldots x_n \leqslant (S/n)^n.$$

Recasting this inequality in terms of the nth elementary symmetric function $E_n(x_1, \ldots, x_n) = x_1 \ldots x_n$ we get

$$E_n(x_1, \ldots, x_n) \leqslant E_n(S/n, \ldots, S/n).$$

In the proof of the birthday inequality, we showed that the elementary symmetric functions $E_k(\mathbf{p})$ are Schur-concave, at least for probability distributions $\mathbf{p}$. In fact, the same proof shows that the elementary symmetric functions are Schur-concave on the set of all n-tuples with nonnegative coordinates; they need not sum to 1. In particular, the function E_n is Schur-concave and the arithmetic-geometric mean inequality is equivalent to the fact that E_n attains its maximum value at $(S/n, \ldots, S/n)$.

Now we return to probability and give another example of a function of a probability vector that is Schur-convex and thus obtains its minimum value at the uniform probability distribution $\mathbf{u}$.

The collector's inequality There are n distinct baseball cards; one of these is included in each package of bubble gum. The collector's goal is to obtain at least one copy of each card. Unfortunately, all selections are made blindly. (You have to buy the bubble gum before you can open the package and see which card is inside.) The collector's problem is this: What is the average (expected) number of packages required in order to obtain a complete collection of all n baseball cards?

First, assume that the cards are uniformly distributed in the gum packages. Then the average number of packages required to collect a complete set is given by,

$$A(\mathbf{u}) = n/n + n/(n-1) + \cdots + n/2 + n/1. \tag{11}$$

This well-known formula appears in Feller's book [2, p. 225]. The proof depends on this intuitive result: If p is the probability of success in a sequence of Bernoulli trials, then the average number of trials required to obtain a success is $1/p$.

But what if the cards are not distributed uniformly in the gum packages? Then the average number of packages $A(\mathbf{p})$ required to collect a complete set depends on the probability vector $\mathbf{p} = (p_1, \ldots, p_n)$, where p_i is the probability that a randomly chosen gum package contains card i. The result we wish to point out is that $A(\mathbf{p})$ is a Schur-convex function of the probability vector $\mathbf{p}$. That is, if $\mathbf{p} \gg \mathbf{q}$, then $A(\mathbf{p}) \geqslant A(\mathbf{q})$. Thus, the minimum value of $A(\mathbf{p})$ occurs when $\mathbf{p}$ is the uniform distribution $\mathbf{u} = (1/n, \ldots, 1/n)$, i.e., when each of the cards is just as likely to occur in a package as any other card. The proof that $A(\mathbf{p})$ is Schur-convex follows from a theorem in Marshall and Olkin [3, pp. 297, 306]. We present a simple proof that $A(\mathbf{p})$ is Schur-convex in the case $n = 2$. In this case there are only two cards to collect, say Aaron and Ruth. Suppose these cards occur with probabilities p_1 and p_2, respectively. If we get Aaron in the first package, then (by the intuitive result above) the expected number of additional packages required to get Ruth is $1/p_2$. So in this case we expect to buy a total of $1 + 1/p_2$ packages. Likewise, if we get Ruth first, then we expect to buy a total of $1 + 1/p_1$ packages. Thus

$$\begin{aligned} A(p_1, p_2) &= p_1(1 + 1/p_2) + p_2(1 + 1/p_1) \\ &= 1 + p_1/p_2 + p_2/p_1. \end{aligned} \tag{12}$$

Equipped with formula (12) for $A(\mathbf{p})$, we proceed to show that $A(\mathbf{p})$ is Schur-convex. This amounts to showing that $A(1 - p_2, p_2)$ is a decreasing function in p_2, for p_2 in the interval $[0, 1/2]$. (By symmetry, we may assume that $0 \leqslant p_2 \leqslant 1/2$.) But

$$\begin{aligned} A(1 - p_2, p_2) &= 1 + (1 - p_2)/p_2 + p_2/(1 - p_2) \\ &= 1/(p_2(1 - p_2)) - 1. \end{aligned}$$

The quadratic $p_2(1-p_2)$ increases on the interval $[0, 1/2]$, so $A(1-p_2, p_2)$ decreases on the same interval. Thus $A(\mathbf{p})$ is Schur-convex for $n = 2$.

We have given two examples of a function in several variables with constant sum that attains its maximum or minimum value at the point where all the variables are equal. When you encounter such an inequality, you should suspect that a stronger version may be true. The function might also be Schur-convex or Schur-concave. For a superb account of the history, theory, and applications of majorization see Marshall and Olkin's definitive work [3].

REFERENCES

1. D. M. Bloom, A birthday problem, *Amer. Math. Monthly* 80 (1973), 1141–2.
2. William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd edition, John Wiley & Sons, New York, 1970.
3. Albert W. Marshall and Ingram Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
4. Frederick Mosteller, *Fifty Challenging Problems in Probability with Solutions*, Addison-Wesley, Reading, MA, 1965.
5. A. G. Munford, A note on the uniformity assumption in the birthday problem, *The Amer. Statist.* 31 (1977), 119.
6. Philip F. Rust, The effect of leap years and seasonal trends on the birthday problem, *The Amer. Statist.* 30 (1976), 197–8.

Monte Carlo Simulation of Infinite Series

FREDERICK SOLOMON
Warren Wilson College
Swannanoa, NC 28778

Let c be a real number. A *Monte Carlo simulation* of c consists of these steps: First, a random experiment is devised and a random variable X is defined for the experiment such that the expectation of X satisfies $c = E(X)$. Second, the random experiment is performed a large number, K of times; let X_i denote the value of X on the ith experiment. By the Strong Law of Large Numbers,

$$c = E(X_i) = \lim_{K \to \infty} \frac{X_1 + X_2 + \cdots + X_K}{K}$$

with probability 1. Consequently, for large K, c "should be" approximately $(X_1 + X_2 + \cdots + X_K)/K$. This ratio is then the Monte Carlo estimate of c.

This method also covers the case in which the constant c is represented as the *probability of an event* A rather than directly as the expectation of a random variable. Namely if $c = P(A)$, then perform the random experiment through which event A is defined and define the random variable X by

$$X = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{if } A \text{ does not occur.} \end{cases}$$