Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and Inference 130 (2005) 377–389

journal of statistical planning and inference

www.elsevier.com/locate/jspi

The matching, birthday and the strong birthday problem: a contemporary review

Anirban DasGupta

Department of Statistics, Purdue University, 1399 Mathematical Science Building, West Lafayette, IN 47907, USA

Received 7 October 2003; received in revised form 1 November 2003; accepted 10 November 2003

Dedicated to Herman Chernoff with appreciation and affection on his 80th birthday

Available online 29 July 2004

Abstract

This article provides a contemporary exposition at a moderately quantitative level of the distribution theory associated with the matching and the birthday problems. A large number of examples, many not well known, are provided to help a reader have a feeling for these questions at an intuitive level.

Keywords: Birthday problem; Coincidences; Matching problem; Poisson; Random permutation; Strong birthday problem

1. Introduction

My first exposure to Professor Chernoff's work was in an asymptotic theory class at the ISI. Later I had the opportunity to read and teach a spectrum of his work on design of experiments, goodness of fit, multivariate analysis and variance inequalities. My own modest work on look-alikes in Section 2.8 here was largely influenced by the now famous Chernoff faces. It is a distinct pleasure to write this article for the special issue in his honor.

This article provides an exposition of some of the major questions related to the matching and the birthday problems. The article is partially historical, and partially forward looking. For example, we address a new problem that we call the *strong birthday problem*. Section 2 takes the birthday problem, and gives a review of the major results in the canonical birthday problem, including the asymptotic Poisson theory, and the case of unequal probabilities. It

E-mail address: dasgupta@stat.purdue.edu (A. DasGupta).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2003.11.015

also discusses how the results change in a Bayesian formulation, with Dirichlet priors on the probabilities of different birthdays. We also introduce a new problem, which we call the strong birthday problem, and discuss an application of it to a problem of interest in criminology and sociology. Section 3 addresses the matching problem, and gives a collection of new and known results on Poisson asymptotics, including some very sharp bounds on the error of the Poisson approximation. It also provides a review of the modern asymptotics on the problem of the longest increasing subsequence and some review of other interesting patterns in random permutations. Feller (1966) is still the best introduction to these charming questions.

2. Birthday and strong birthday problems

The classical birthday problem asks, what is the probability of finding at least one similar pair having the same birthday in a group of n individuals. This problem was initiated by von Mises in 1932. The strong birthday problem asks, what is the probability that each one in a group of n individuals is a member of some similar pair. Another way to ask the same question is what is the probability that everyone in a group of n individuals has a birthday shared by someone else in the group. In the classical birthday problem, the smallest n for which the probability of finding at least one similar pair is greater than .5 is $n = 23$. In the strong birthday problem, the smallest n for which the probability is more than .5 that everyone has a shared birthday is $n = 3064$. The latter fact is not well known. We will discuss the canonical birthday problem and its various variants, as well as the strong birthday problem in this section.

2.1. *The canonical birthday problem*

Let I_{ij} be the indicator of the event that individuals i, j have the same birthday. Then, the number of similar pairs, is $W = \sum_{i<j} I_{ij}$. The I_{ij} are not independent. Thus the exact distribution of W is complicated, even for the case of all days being equally likely to be the birthday of any given individual. However, the question originally raised by von Mises is answered easily. Indeed, $p_n = P(\text{at least one similar pair}) = 1 - P(\text{no similar pair}) = 1 - \frac{\prod_{i=1}^{n-1}(365-i)}{365^{n-1}}$, discounting leap years, and making the equally likely and independence among individuals assumptions. The probability of at least one similar pair is as follows for various values of n.

Example 1. Thus, in a gathering of 50 people, it is highly probable that at least one similar pair will be found. As remarked before, the exact distribution of W, the total number of similar pairs is too complicated. Consider the more general case of m equally likely birthdays. Then, the distribution of W can be well approximated by a Poisson distribution.

Table 1
P(at least one similar pair)

n	2	3	4	5	10	20	23	30	50
p_n	.003	.008	.016	.027	.117	.411	.507	.706	.970

The Poisson approximation to the probability of at least one similar pair yields very good results.

The following theorem can be seen in various places, including Arratia et al. (1989, 1990), Stein (1986), Barbour et al. (1992), and Diaconis and Holmes (2002).

Theorem 1. *If $m, n \to \infty$ in such a way that $\frac{n(n-1)}{m} \to 2\lambda$, then $W \overset{\mathscr{L}}{\Rightarrow} \mathrm{Poi}(\lambda)$.*

If m, n are 365 and 23, respectively, then taking $2\lambda = \frac{22\times 23}{365} = 1.3863$, and using the Poisson approximation we get, $P(W \geqslant 1) = 1 - \mathrm{e}^{-.69315} = .500001$. This is a good approximation to the true value of .507 (Table 1).

A slightly different question is the following: if we interview people, one by one, until we find a similar pair, how many should be interviewed? If we denote this number by N, then $E(N) = \sum_{n=0}^{m}(1-p_n)$. Calculation using the formula above for p_n gives $E(N) = 24.6166$. The variance of N equals 148.64.

2.2. Matched couples

An interesting variation of the canonical birthday problem is the following question: suppose n couples are invited to a party. How surprised should one be if there are at least two husband–wife pairs such that the husbands have the same birthdays and so do their wives? The answer is that one should be considerably surprised to observe this in normal gatherings. In fact, changing m to m^2 in the canonical birthday problem, the probability of finding no matched couples is $\prod_{i=1}^{n-1}(1 - \frac{i}{m^2})$. With $m = 365$, this is .9635 if there are $n = 100$ couples in the party. The probability of no matched couples falls below 50% for the first time when $n = 431$.

2.3. Near hits

Abramson and Moser (1970) discuss the case of *nearly the same birthdays*. Thus, one can ask what is the probability of finding at least one pair in a group of n people with birthdays within k calendar days of each other's? From the point of view of coincidences, the case $k = 1$ may be the most interesting.

Let $p(m, n, k)$ denote the probability that in a group of n people, at least one pair with birthdays within k days of each other's exists, if there are m equally likely birthdays. Abramson and Moser (1970) show that

$$p(m, n, k) = \frac{(m - nk - 1)!}{(m - n(k+1))!\,m^{n-1}}. \tag{1}$$

Example 2. Using formula (1), the probability of finding a pair of people with birthdays within one calendar day of each other's is .08 in a group of five people, .315 in a group of 10 people, .483 in a group of 13 people, .537 for 14 people, .804 for 20 people, and .888 for 23 people. A quite accurate approximation to the smallest n for which $p(m, n, k)$ is .5 is $n = 1.2\sqrt{\frac{m}{2k+1}}$; see Diaconis and Mosteller (1989).

2.4. Similar triplets and p of a kind

The answer $n=23$, in the canonical birthday problem is surprising. People do not expect that a similar pair would be likely in such a small group. It turns out that while a similar pair is relatively easy to find, similar triplets are much harder to observe.

As before, one can ask for the distribution of $W=$ number of similar triplets, and the smallest n such that $P(W\geqslant 1)\geqslant .5$. Or one can define N_p as the minimum number of people one needs to interview before p people with a common birthday have been found; $p=3$ corresponds to a similar triplet.

Using multinomial probabilities, one can write a messy expression for $P(W\geqslant 1)$:

$$P(W\geqslant 1)=1-\sum_{i=0}^{[n/2]}\frac{m!n!}{i!(n-2i)!(m-n+i)!2^i m^n}. \tag{2}$$

Example 3. If $m=365$, $P(W\geqslant 1)=.013$ for 23 people, .067 for 40 people, .207 for 60 people, .361 for 75 people, .499 for 87 people, .511 for 88 people, and .952 for 145 people. These numbers show how much harder it is to find a similar triplet compared to a similar pair.

A first-order asymptotic expression for $E(N_p)$ is given in Klamkin and Newman (1967). They show that

$$E(N_p)\sim(p!)^{1/p}\Gamma\left(1+\frac{1}{p}\right)m^{1-1/p} \tag{3}$$

for fixed p, and as $m\to\infty$. For $p=3$, the asymptotic expression gives the value 82.87. We suspect that it is not too accurate, as the exact value is about 88.

2.5. Unequal probabilities and Bayesian versions

Diaconis and Holmes (2002) give results on Bayesian versions of the canonical birthday problem. The vector of probabilities $(p_1, p_2, \ldots, p_m)$ of the m birthdays is unknown, and a prior distribution on the $(m-1)$ dimensional simplex Δ_m is assumed. The questions of interest are: The marginal probability (i.e., integrated over the prior) of finding at least one similar pair for fixed n (the group size), and the limiting distribution (if one exists) of the total number of distinct similar pairs when $n\to\infty$.

If the vector of cell probabilities has an exchangeable Dirichlet$(\alpha, \alpha, \ldots, \alpha)$ prior, then $P(W\geqslant 1)=1-\prod_{i=1}^{n-1}\frac{\alpha(m-i)}{m\alpha+i}$. This can be derived by direct integration or as Diaconis and Holmes (2002) derive by embedding it into the Polya urn scheme.

Since there is an exact formula, the smallest n required for the marginal probability $P(W\geqslant 1)$ to be $\geqslant .5$ can be calculated easily. The table below shows the required n as a function of α. $\alpha=1$ corresponds to the uniform prior on the simplex; large α corresponds approximately to the classical case, i.e., a point prior. Notice that a smaller n suffices in the Bayesian version. This is because the exchangeable Dirichlet priors allow some clumping (in terms of distribution of the people into the various birthdays) and so a similar pair is more likely in the Bayesian version.

α	.5	1	2	3	4	5	20
n	14	17	19	20	21	21	23

The Poisson limit theorem in Diaconis and Holmes (2002) explains more clearly why similar pairs are more likely to be found in the Bayesian version. The theorem says the following.

Theorem 2. *Suppose* $m, n \to \infty$ *in such a way that* $\frac{n(n-1)}{m} \to 2\lambda$. *Then, under the exchangeable* Dirichlet$(\alpha, \alpha, \ldots, \alpha)$ *prior,* $W \overset{\mathscr{L}}{\Rightarrow} \text{Poi}(\frac{\alpha+1}{\alpha}\lambda)$.

Thus, under the uniform prior, one would expect about twice as many similar pairs as in the classical case. This is very interesting. For other references on the unequal probability case, Camarri and Pitman (2000) is an excellent source.

2.6. Strong birthday problem

The “strong birthday problem” asks for the probability of a much stronger coincidence than does the canonical birthday problem. It asks what is the probability in a group of n people that everyone in the group shares his or her birthday with someone else in the group? If we let the number of unique people be N, then the problem is to give a formula for $P(N = k)$. The strong birthday problem has applications to the interesting problem of *look-alikes*, which is of interest to criminologists and sociologists. The material in this section is taken from DasGupta (2001).

Using our earlier notation, in the equally likely and independent case, writing $S_i = \frac{m!n!(m-i)^{n-i}}{i!(m-i)!(n-i)!m^n}$, we obtain

$$P(N = k) = \sum_{i=k}^{n} (-1)^{i-k} \frac{i!}{k!(i-k)!} S_i. \tag{4}$$

This is a consequence of standard formulae for the probability of occurrence of k out of n events. Using $m = 365$ and $k = 0$, one can compute the probability p_n that everyone in a group of n individuals has a shared birthday.

Example 4. Thus, in a group of 3064 people, the probability is greater than half that everyone has a shared birthday (Table 2).

An accurate iterative approximation of the smallest n required to make $p_n = p$ is $\frac{n_1}{m} = \log\left(\frac{m}{\log\frac{1}{p}}\right)$, $\frac{n_i}{m} = \frac{n_1}{m} + \log(\frac{n_i - 1}{m})$. For $p = .5$, five iterations yield the value 3064. Thus, five iterations give the correct value of n.

Table 2
P(each person has a shared birthday)

n	p_n
2000	.0001
2500	.0678
2700	.1887
2800	.2696
3000	.4458
3063	.4999
3064	.5008
3500	.7883
4000	.9334
4400	.9751

Under certain configurations of m and n, the number of unique individuals has a Poisson limit distribution. Under *other* configurations, there can be other limiting distributions. By linking N to the number of cells with exactly one ball in a multinomial allocation, the various limiting distributions corresponding to various configurations of m, n can be obtained from the results in Kolchin et al. (1978). We will report only the limiting Poisson case as that is the most interesting one.

Theorem 3. *Suppose* $\frac{n}{m} = \log n + c + \mathrm{o}(1)$. *Then, under the equal probability and independence* (*of the people*) *assumptions,* $N \overset{\mathscr{L}}{\Rightarrow} \mathrm{Poi}(\mathrm{e}^{-c})$.

For example, if $m = 365$ and $n = 3064$, then using $c = \frac{3064}{365} - \log 3064 = .367044$, one gets the Poisson approximation $P(N = 0) = \mathrm{e}^{-\mathrm{e}^{-.367044}} = .50018$, a remarkably good approximation to the exact value .50077.

2.7. Bayesian versions

For general arbitrary birthday probabilities $p_1, p_2, \ldots, p_m$, the distribution of N can become very complicated. Of course, the mean and the variance are easy to find. We have

$$E(N) = n \sum_{k=1}^{m} p_k (1 - p_k)^{n-1}$$

and

$$\mathrm{Var}(N) = n(n-1) \sum_{i \neq j} p_i p_j (1 - p_i - p_j)^{n-2} + E(N) - (E(N))^2. \tag{5}$$

If we let the vector of cell probabilities have an exchangeable Dirichlet$(\alpha, \alpha, \ldots, \alpha)$ prior distribution, then the marginal expectation of N is found to be

$$E_\alpha(N) = \frac{mn\alpha\Gamma(m\alpha)\Gamma(\alpha(m-1)+n-1)}{\Gamma((m-1)\alpha)\Gamma(m\alpha+n)}. \tag{6}$$

For a fixed vector of cell probabilities $p_1, p_2, \dots, p_m$,

$$P(N=0)=1-\sum_{j=1}^{\min(m,n)}(-1)^{i-1}\frac{n!}{j!(n-j)!}\times\sum_{i_1\neq i_2\neq\cdots\neq i_j}p_{i_1}p_{i_2}\dots p_{i_j}(1-p_{i_1}-p_{i_2}-\cdots-p_{i_j})^{n-j}.$$

This can be integrated with respect to a Dirichlet$(\alpha, \alpha, \dots, \alpha)$ density by using the Liouville integral formula $\int_{\Delta_s} p_1^{a_1}p_2^{a_2}\dots p_s^{a_s} f(p_1+p_2+\cdots+p_s)\,\mathrm{d}p_1\dots\mathrm{d}p_s=\frac{\Gamma(1+a_1)\dots\Gamma(1+a_s)}{\Gamma(s+a_1+\cdots+a_s)}\cdot\int_0^1 x^{s+a_1+\cdots+a_s-1}f(x)\,\mathrm{d}x$. The resulting expressions are still closed form, but messy.

For $m<n$, and the uniform prior, this works out to be

$$P_u(N=0)=1-\frac{(m-1)!m!n!}{(m+n-1)!}\sum_{j=1}^{m-1}(-1)^{j-1}\frac{(m+n-2j-1)!}{j!(n-j)!(m-j-1)!(m-j)!}. \quad (7)$$

However, for large values of m, n, due to the alternating signs in the sum in formula (7), the computation appears to turn unreliable. Thus, asymptotics would be useful again. From formula (7), one can show that if $n\sim\theta m^2$, then $P(N=0)\to \mathrm{e}^{-1/\theta}$. This gives $n\approx 192,000$ as the required value of n for the Bayesian probability that everyone has a shared birthday to be 50%. On comparing this to the value $n=3064$ when all birthdays are assumed equally likely, we see that a *much* larger group is needed in the Bayesian version of the problem. This is because the uniform prior on the simplex allows the probability of $N=0$ to be small over a large part of the simplex. Thus, the canonical birthday problem and the strong birthday problem behave differently in the Bayesian formulation of the problem.

2.8. Eyewitness testimony and look-alikes

In criminal investigations, law enforcement often circulates a picture of a suspect drawn on the basis of information provided by a witness on some key physical features. Instances of erroneous apprehension are common, because an innocent person happens to look like the person drawn in the picture. The various configurations of physical features can be regarded as the cells of a multinomial and people regarded as balls. Thus, if we were to consider 10 physical features, each with three different categories (such as tall-average-short for height), then we have a multinomial with $m=3^{10}$ cells. This is a huge number of cells. Yet, if n, the relevant population size, is large enough, then the number of cells with 2 or more balls would be large too. This would imply that the person implied in the picture may have a *look-alike*. Thus, the calculations in the strong birthday problem have application to criminology, in particular, assessing the likelihood of misapprehension in criminal incidents.

Example 5. Using a set of 14 physical features, such as sex, height, heaviness, facial shape, nasal elevation, size of pupils, eyebrow thickness, size of head, etc., with 2–4 divisions within each feature, with a total of 1.12 million cells in all, and making the (unrealistic) assumption that all cells are equally likely, we found that in the city of New York (assuming a population size of $n=8$ millions), the number of people *without* a look-alike is approximately distributed as Poisson with a mean of about 6300. This is a consequence of general asymptotic theory on number of cells with exactly one ball in an equiprobable multinomial allocation; see Kolchin et al. (1978). A realistic probabilistic analysis would be difficult here because we have no way to ascertain reliable values for the probabilities of the 1.12 million cells of physical configurations. According to the equiprobable scheme, it would not be surprising at all to find a look-alike of almost anyone living in New York city. Better models for this problem which can be analyzed should be of interest to criminologists.

3. Matching problems

Card matching using two decks is the most common form of what are collectively known as matching problems. Imagine one deck of 52 cards labeled as $1, 2, \ldots, 52$ and another deck shuffled at random. Pick the cards of the shuffled deck one at a time and if for a given i, the card picked is numbered i, then its position in the shuffled deck matches its position in the unshuffled deck. The basic matching problem asks questions about the total number of matches, such as the expected value, and its distribution. The problem goes back to at least Montmort in the early 18th century. Many surprising results are known about the total number of matches. The problem is often stated in other forms, such as returning hats or stuffing letters at random. The mathematics of the basic matching problem uses the elegant combinatorics of *random permutations*.

3.1. Unrestricted matching problem

The problem can be formulated as follows. Let π denote a random permutation of the n numbers $1, 2, \ldots, n$; i.e., the probability that π is any of the $n!$ permutations of $\{1, 2, \ldots, n\}$ is $\frac{1}{n!}$. We call the number i a fixed point of π if $\pi(i) = i$. Denote by Z the total number of fixed points of π. Obviously, Z can take the values $0, 1, \ldots, n$.

One can write the exact distribution of Z for any n. By standard combinatorial arguments, $P(Z = k) = \frac{1}{k!}\sum_{i=0}^{n-k}(-1)^i\frac{1}{i!}$. In particular, the probability of no matches, $P(Z = 0) = \sum_{i=0}^{n}\frac{(-1)^i}{i!}$; this converges, as $n \to \infty$ to $\mathrm{e}^{-1} \approx .36788$. The convergence is extremely rapid, and with two decks of 52 cards each, the probability of no matches is almost identical to the limiting value .36788. Note that e^{-1} is the probability that a Poisson random variable with mean 1 assumes the value zero. Noting that the exact mean of Z is *always* 1, this might lead us to suspect that Z has a limiting Poisson distribution with mean 1. This is true, and in fact many strong results on the convergence are known. For purpose of illustration, first we provide an example.

Example 6. Consider the distribution of Z when $n=10$. By using the formula given above, the distribution is as follows (Table 3).

The rest of the probabilities are too small. The most striking aspect of Table 3 is the remarkable accuracy of the Poisson approximation. The other intriguing thing one notices is that the Poisson approximation appears to alternately under and overestimate $P(Z = k)$ for successive values of k for $k \geqslant 4$. It is interesting that these empirical phenomena are

Table 3
Accuracy of Poisson approximation

k	$P(Z=k)$	Poisson approx.
0	.36788	.36788
1	.36788	.36788
2	.18394	.18394
3	.06131	.06131
4	.01534	.01533
5	.00306	.00307
6	.00052	.00051

in fact analytically provable. See DasGupta (1999) and also consult Diaconis and Holmes (2002).

3.2. Error of Poisson approximation

Theorem 4. (a) *Let* $d_{TV} = \frac{1}{2}\sum_{k=0}^{\infty} |P(Z=k) - P(\text{Poisson}(1)=k)|$. *Then* d_{TV} *admits the integral representation*

$$d_{TV} = \frac{1}{n!}\int_0^1 \mathrm{e}^{-t} t^n \, \mathrm{d}t + \frac{1}{(n-1)!}\int_0^1 (2-t)^{n-1} t^{-n-1} \gamma(n+1, t)\, \mathrm{d}t, \tag{8}$$

where $\gamma(n+1, t)$ *denotes the incomplete Gamma function* $\int_0^t \mathrm{e}^{-x} x^n \, \mathrm{d}x$.

(b) $\quad d_{TV} \leqslant \dfrac{2^n}{(n+1)!} \quad$ *for every* $n \geqslant 2$.

The integral representation in (a) implies the error bound in (b). The error bound explains why the Poisson approximation in Table 3 is so sharp. Although much easier proofs can be given, the error bound also implies that $Z \overset{\mathscr{L}}{\Rightarrow}$ Poisson(1) as $n \to \infty$. It should be added that error bounds on the Poisson approximation can also be found by the coupling method of Stein-Chen (see Arratia et al. (1990) for a lucid exposition), but they are not as strong as the more direct bound in part (b) above.

The sign-change property of the error in the Poisson approximation is stated next.

Theorem 5. (a) *For n even,* $P(Z=k) < (>) P(\text{Poisson}(1)=k)$ *according as* $k<n$ *is odd or even; the opposite inequalities hold for n odd.*

(b) *The inequalities of part* (a) *hold when* $(Z=k)$, $(\text{Poisson}(1)=k)$ *are replaced by* $(Z \leqslant k)$, $(\text{Poisson}(1) \leqslant k)$.

Another fascinating fact about the matching problem is that the first n moments of Z and of the Poisson(1) distribution exactly coincide! This provides another proof of Z having a limiting Poisson(1) distribution. How about the subsequent moments? They do not coincide. In fact, the difference diverges.

Table 4
P(no matches)

Expected deck size	Uniform deck	Geometric deck
5	.315	.435
25	.357	.382
50	.363	.375

3.3. Random deck size

How is the total number of matches distributed if the size of the two decks (which we assume to be equal) is random? Under certain assumptions on the size of the deck, the convergence of the total number of matches to the Poisson(1) distribution is still true, but it need not be true in general. For geometric decks with empty decks allowed, a very neat result holds.

Theorem 6. *Suppose the size N of the deck is distributed as* Geometric(p), *with mass function* $p(1-p)^n$, $n \geqslant 0$. *Then the* (*marginal*) *distribution of Z is a Poisson, with mean* $1-p$.

If p is parameterized by m, with $p_m \to 0$, then, still, $Z \overset{\mathscr{L}}{\Rightarrow} \text{Poisson}(1)$.

How does the probability of at least one match behave for random decks? For geometric decks, matches get less likely, as is evident from Theorem 6. But, interestingly, for certain other types of random decks, such as uniform or Poisson decks, matches become *more* likely than the nonrandom case. Here, is an illustrative example (Table 4).

Example 7.

3.4. Near hits

Suppose a person claiming to have psychic powers is asked to predict the numbers on the cards in a shuffled deck. If the person always predicts the number on the card correctly or misses the correct number by 1, how surprised should we feel? Thus, if π is a random permutation of $\{1, 2, \ldots, n\}$, what is the probability that $|\pi(i) - i| \leqslant 1$ for every $i = 1, 2, \ldots, n$? More generally, one can ask what is $P(\max_{1 \leqslant i \leqslant n} |\pi(i) - i| \leqslant r)$, where $r \geqslant 0$ is a fixed integer?

Example 8. For $r = 0, 1, 2$ a relatively simple description can be given. Of course, for $r = 0$, the probability is $\frac{1}{n!}$, and thus even with $n = 5$, one should be considerably surprised. If $r = 1$, the probability is $\frac{F_{n+1}}{n!}$ where F_n is the nth Fibonacci number. This works out to 1, .5, .208, .067, .018 and .004 for $n = 2, 3, 4, 5, 6, 7$, respectively. Thus, if someone was able to call the numbers within an error of 1, we should be considerably surprised even when n is just 6. How about $r = 2$? In this case, the probabilities work out to 1, 1, .583, .258, .101,

.034, .01 and .0026 for $n = 2, 3, \ldots, 9$ respectively. Thus, calling the numbers within an error of 2 should be of considerable surprise if n is 8.

See Tomescu (1985) for these results.

3.5. *Longest increasing subsequence*

Consider a random permutation of $\{1, 2, \ldots, n\}$. Should we be surprised if we see a long increasing (or decreasing) subsequence? To answer this question with precision, one would need to have information about the distribution and asymptotic growth rate of the length of the longest increasing subsequence.

It has been known for a long time that a monotone sequence of length of the order of$\sqrt{n}$ always exists for any real sequence of length n (Erdös and Szekeres (1935)); actually Erdös and Szekeres prove a more general result. One may suspect because of this that the length of the longest increasing subsequence of a random permutation grows asymptotically at the rate$\sqrt{n}$; see Ulam (1961). But an actual proof, for example, a proof of the existence of a weak limit or a proof that the expected length grows at the$\sqrt{n}$ rate involve intricate arguments.

Thus, let I_n denote the length of the longest increasing subsequence of a random permutation π of $\{1, 2, \ldots, n\}$. Then, $\frac{I_n}{\sqrt{n}}$ converges in probability to 2, and furthermore, $\frac{E(I_n)}{\sqrt{n}} \to 2$. In fact, even second order asymptotics for $E(I_n)$ are known; settling a longstanding conjecture founded on Monte Carlo and other evidence, Baik et al. (1999) established the result $\frac{E(I_n)-2\sqrt{n}}{n^{1/6}} \to c_0$, where c_0 is the mean of the *Tracy–Widom distribution* on the reals. An approximate numerical value for c_0 is -1.7711. The CDF of the Tracy–Widom distribution does not have closed form formula, but numerical evaluation is possible, by numerical solution of a corresponding differential equation. In fact, one has the remarkable result that $\frac{(I_n-2\sqrt{n})}{n^{1/6}} \overset{\mathscr{L}}{\Rightarrow} L$, L having the Tracy–Widom distribution. See Tracy and Widom (1994), Baik et al. (1999) and Aldous and Diaconis (1999) for these results. A very readable review of these results is available in the Aldous and Diaconis (1999) reference.

It is also possible to describe, for each fixed n, the distribution of I_n by linking it to a suitable distribution on the possible shapes of a *Young tableaux*. Evolution of these results can be seen in Hammersley (1972), Baer and Brock (1968), Logan and Shepp (1977), and Versik and Kerov (1977). It is also true that for ‘most’ random permutations, the length of the longest increasing subsequence stays ‘close’ to the $2\sqrt{n}$ value. Precise statements in terms of large deviations can be seen in Talagrand (1995), Steele (1997) and the references mentioned above.

Computing the actual value of the length of the longest increasing subsequence of a given permutation is an interesting problem, and there is substantial literature on writing efficient algorithms for this problem. The interested reader should consult Steele (1995) for a survey.

3.6. *Surprise in seeing other patterns*

Numerous other interesting patterns in sequence matching have been discussed in the literature. We will briefly discuss the case of *falls, and up-down permutations* and the surprise factor associated with each one.

Table 5
P(an up-down permutation)

n	
2	.5
3	.333
4	.208
5	.133
7	.054
10	.014
15	.001
20	.00015

A permutation π of $\{1, 2, \ldots, n\}$ has a *fall* at location i if $\pi(i+1) < \pi(i)$, with the convention that the last location n is always counted as a fall. Seeing about how many falls should surprise us? Surprisingly, for every fixed n, the distribution of the total number of falls can be explicitly described. It is related to the sequence of Eulerian numbers $A(n,k) = (n+1)!\sum_{i=0}^{k-1} \frac{(-1)^i}{i!(n-i+1)!}(k-i)^n$ (not to be confused with Euler numbers); see Blom et al. (1991). Denoting the number of falls in a random permutation by N_f, $P(N_f = k) = \frac{A(n,k)}{n!}$. Calculation using the Eulerian numbers shows that seeing 4 falls when $n = 6$, 5 falls when $n = 7$, and 6 falls when $n = 8$ would not be much of a surprise. Of course, the expected number of falls is $\frac{n+1}{2}$.

Example 9. In terms of permutations with structures, up-down permutations are among the ones that should surprise an observer mostly.

A permutation π is called an up-down permutation if $\pi(1) < \pi(2) > \pi(3) < \pi(4) > \cdots$; obviously, such a permutation is extremely patterned and one would feel surprised to see it. If u_n denotes the number of up-down permutations of $\{1, 2, \ldots, n\}$, then the exponential generating function of the sequence u_n, i.e., $G_n(t) = \sum_{n=0}^{\infty} \frac{u_n t^n}{n!}$ equals $\sec t + \tan t$; see Tomescu (1985). Thus, $u_n = \frac{\mathrm{d}^n}{\mathrm{d}t^n}(\sec t + \tan t)|_{t=0}$. For example, $u_5 = 16$, and $u_{10} = 50\,521$. The probability of seeing an up-down permutation is listed in the Table 5 for some selected values of n; it would be an extreme surprise to observe one if n was 15 or so.

Acknowledgements

It is a pleasure to thank Jim Pitman and Shelley Zacks for their help and comments.

References

Abramson, M., Moser, W., 1970. More birthday surprises. Amer. Math. Monthly 7 (7), 856–858.

Aldous, D., Diaconis, P., 1999. Longest increasing subsequences: from patience sorting to the Baik–Deift–Johansson theorem. Bull. Amer. Math. Soc. 36 (4), 413–432.

Arratia, R., Goldstein, L., Gordon, L., 1989. Two moments suffice for the Poisson approximation: the Chen–Stein method. Ann. Probab. 17 (1), 9–25.

Arratia, R., Goldstein, L., Gordon, L., 1990. Poisson approximation and the Chen–Stein method. Statist. Sci. 5, 403–434.

Baer, R.M., Brock, P., 1968. Natural sorting over permutation spaces. Math. Comput. 22, 385–410.

Baik, J., Deift, P., Johansson, K., 1999. On the distribution of the length of the longest increasing subsequence of random permutations. J. Amer. Math. Soc. 1 (2), 1119–1178.

Barbour, A.D., Holst, L., Janson, S., 1992. Poisson Approximation. Clarendon Press, Oxford University Press, New York.

Blom, G., Holst, L., Sandell, D., 1991. Problems and Snapshots from the World of Probability. Springer, New York.

Camarri, M., Pitman, J., 2000. Limit distributions and random trees derived from the birthday problem with unequal probabilities. Electron. J. Probab. 5, 1–18.

DasGupta, A., 1999. The matching problem with random decks and the Poisson approximation. Technical report, Purdue University.

DasGupta, A., 2001. Strong birthday problems and look-alikes. Preprint, Purdue University.

Diaconis, P., Mosteller, F., 1989. Methods for studying coincidences. J. Amer. Statist. Assoc. 8 (4), 408, 853–861.

Diaconis, P., Holmes, S., 2002. A Bayesian peek at Feller, vol. I. Sankhýa, Ser. A, Special Issue in Memory of D. Basu 64 (3(ii)), 820–841.

Erdös, P., Szekeres, G., 1935. A combinatorial theorem in geometry. Comput. Math. 2, 463–470.

Feller, W., 1966. An Introduction to Probability Theory and its Application, vol. I. Wiley, New York.

Hammersley, J.M., 1972. A few seedlings of research. Proc. Sixth Berkeley Symposium on Mathematics Statistics, and Probability, vol. I, University of California Press, California, pp. 345–394.

Klamkin, M.S., Newman, D.J., 1967. Extensions of the birthday surprise. J. Combin. Theory 3, 279–282.

Kolchin, V.F., Sevast'yanov, B.A., Chistyakov, V.P., 1978. Random Allocations. Wiley, New York.

Logan, B.F., Shepp, L., 1977. A variational problem for random Young tableaux. Adv. Math. 26, 206–222.

Stein, C., 1986. Approximate Computation of Expectations. IMS Monograph Series. Hayward, CA.

Steele, J.M., 1995. Variations on the long increasing subsequence theme of Erdös and Szekeres. in: Aldous, D., Diaconis, P., Steele, J.M. (Eds.), Discr. Prob. and Algorithms. Springer, New York.

Steele, J.M., 1997. Probability Theory and Combinatorial Optimization. SIAM, Philadelphia.

Talagrand, M., 1995. Concentration of measure and isoperimetric inequalities in product spaces. Publ. Math. IHES 81, 73–205.

Tomescu, I., 1985. Problems in Combinatorics and Graph Theory. Wiley, New York.

Tracy, C.A., Widom, H., 1994. Level-spacing distributions and the Airy kernel. Comm. Math. Phys. 159, 151–174.

Ulam, S., 1961. Monte Carlo calculations in problems of mathematical Physics. in: Beckenbach, E.F. (Ed.), Modern Mathematics for Engineers. McGraw-Hill, New York.

Versik, A.M., Kerov, S., 1977. Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tables. Soviet Math. Dokl. 18, 527–531.