

Sequences, Patterns and Coincidences

Anirban DasGupta

Purdue University, West Lafayette, IN

March 12, 2004

ABSTRACT

This article provides a contemporary exposition at a moderately quantitative level of the distribution theory associated with sequences and patterns in iid multinomial trials, the birthday problem, and the matching problem. The section on patterns includes the classic distribution theory for waiting time for runs and more general patterns, and their means and moments. It also includes the central limit theorem and a.s. properties of the longest run, and modern applications to DNA sequence alignment. The section on birthdays reviews the Poisson approximation in the classical birthday problem, with new bounds on the total variation distance. It also includes a number of variants of the classic birthday problem, including the case of unequal probabilities, similar triplets, and the Bayesian version with Dirichlet priors on birthday probabilities. A new problem called the strong birthday problem with application to criminology is introduced. The section on matching covers the Poisson asymptotics, errors of Poisson approximations, and some variants such as near matches. It also briefly reviews the recent extraordinary developments in the area of longest increasing subsequences in random permutations.

The article provides a large number of examples, many not well known, to help a reader have a feeling for these questions at an intuitive level.

Key words : Patterns, Runs, Waiting times, Generating functions, Longest runs, Birthday problem, Poisson distribution, Strong

birthday problem, Prior, Random permutation, Matching, Longest increasing subsequence.

1 Introduction

This article provides an exposition of some of the most common examples and anecdotes that we have come to know as patterns and coincidences. The article is partially historical because it is written for an Encyclopedia. However, the article is also partially forward looking in its choice of examples and topics. Our selection of DNA sequence alignments as a topic is an attempt to make this exposition contemporary. There are other topics that we discuss that have a similar forward looking value.

Coincidences are hard to define. But in our day to day lives, we have all been surprised to see something happen that the mind perceived as accidental synchronicity of two or more events. We think of such an experience as a coincidence. We meet someone at a party and discover that we have the same birthday; we think of it as a coincidence. We go on vacation to a foreign country and meet someone we had not seen for 10 years; we are surprised and think of it as a stunning coincidence. We play a dice game and have three sixes show up in a row; again, we feel amazed and think of it as a coincidence. We read in the newspaper that someone in a small town in New Jersey has won the state lotto two times in four months. We find it unbelievable and think of it as a fantastic coincidence. Diaconis and Mosteller (1989) write :“Coincidences abound in everyday life. They delight, confound, and amaze us. They are disturbing and annoying. Coincidences can point to new discoveries. They can alter the course of our lives.” It is the ‘amazement’ or the ‘surprise’ component that one can attempt to study mathematically. Was what the mind perceived as extraordinarily surprising really surprising ? Can it be explained mathematically ? To explain an observed phenomenon mathematically, one would have to have a mathematical model or structure within which the phenomenon would be analyzed. Many examples of coincidences have natural mathematical models for them. Others do not. We will provide an exposition of what is known in examples where a model is known, and raise new questions and suggest models where a model has not been formulated. In this sense, our article makes an attempt to be partially historical and partially forward looking.

Coincidences can and have been studied from other viewpoints. Einstein was considerably intrigued that gravity and inertia exactly balance each other out in the universe. There is classic literature on study of coincidences from the points of view of psychology, difference between reality and perception, selective memory, and hidden causes. Diaconis and Mosteller (1989) provide a very lucid exposition to that part of the literature and provide many references. We will not go into a review of that literature here.

The article takes one specific example or problem in a section and provides an exposition of that particular problem in that section. We provide many examples and numerical illustrations.

2 An Example : The Next Seat Passenger on an Airplane

Example 1 In her engrossing book *Mind Over Matter*(2003), Karen Christine Cole asks “... Consider two strangers sitting in adjacent seats on an airplane. What are the chances they will have something in common ?” Our first example investigates this interesting question at a mildly mathematical level.

We include only those characteristics such that a coincidence with the next seat passenger would actually cause some surprise. For example, if two men travelling in adjacent seats had wives with the same first name, it would probably cause both of them a chuckle and a surprise. But if discussions turned to politics, and both turned out to be Republicans, it would hardly be a surprise. After all, there are only two or three parties in our political system.

Obviously we have to choose our characteristics. We choose the following : first name, last name, spouse’s first name, father’s name, mother’s name, birthday, spouse’s birthday, father’s birthday, mother’s birthday, profession, spouses’s profession, state of birth, state of residence, spouses’s state of birth, age, university attended, university attended by spouse, favorite actor, favorite

actress, favorite singer, favorite writer, favorite cuisine, hobby, spouses's hobby; a total of 24 characteristics.

For first as well as last names, we take 200 as the number of possibilities; for the others, we take 365 possibilities for the birth-days, 50 for professions, 50 for state of birth and residence, 50 for age(say 20 to 70), 100 for university attended for passenger as well as spouse, 20 for favorite actor and actress, 50 for singer and writer, 10 for favorite cuisine, and 50 for the hobby of the passenger as well as spouse. For the characteristics $1, 2, \dots, k = 24$, denote the number of possibilities allowed by m_1, m_2, \dots, m_k respectively; e.g., $m_1 = 200$.

Defining I_i to be the indicator of whether there is a match in the i th characteristic, the total number of coincidences is simply the sum $T = \sum_{i=1}^k I_i$. We assume the indicator variables I_i to be independent (there is actually some dependence) and the possibilities within a category to be equally likely. Then, the probability of a coincidence with the next seat passenger on at least one of the characteristics equals $1 - \prod_{i=1}^k (1 - \frac{1}{m_i}) = .41$. The probability of a coincidence on two or more characteristics is .09. Computing the probability of three or more coincidences exactly would be a bit tedious. But one can approximate this probability by means of a *Poisson approximation*, and even make statements about the error of the Poisson approximation. Le Cam(1960) (using 2 as the best constant available right now in Le Cam's error bound) gives a simple error bound.

To finish this example, suppose now our traveller is a seasoned one, making one round trip every month. For such a traveller, the chances are more than 50% that on at least one flight in a four month time period, s/he will have a coincidence in two or more of the 24 characteristics with his or her next seat passenger. The chances are 90% that within a year, s/he will have a coincidence in two or more characteristics with his or her next seat passenger. But what would the next seat passenger think if someone actually asked 'what is your mother's name ?'

3 Runs, Sequences and Patterns in Multinomial Trials

3.1 Waiting Time for Success Runs

Example 2 Suppose we toss a fair coin repeatedly until three heads are obtained in a row (generally called a first run of three heads). How many tosses would it take to see a run of three heads ? In fact, let us treat the general case of a coin with probability p for heads in any single toss, writing q for $1 - p$.

Clearly, we cannot see a run of three heads before we have tossed the coin at least three times. If we let N denote the number of tosses it would take to see our run of three heads for the first time, then $p_k = P(N = k)$ satisfies :

$p_3 = p^3; p_4 = qp^3; p_5 = q^2p^3 + pqp^3$, and in general, the three-term recursion $p_k = qp_{k-1} + pqp_{k-2} + p^2qp_{k-3}$. This is adequate to produce the numerical value of any p_k , but we can use the three term recursion to write down a few things analytically. Indeed, such a recursion is *the key step* in many of these problems on patterns and coincidences.

Thus, if $\psi(t) = E(t^N)$ denotes the generating function of the p_n sequence, then a little algebraic manipulation of the three term recursion produces the closed form formula :

$$\psi(t) = \frac{p^3t^3(1-pt)}{1-t+qp^3t^4} \quad (1)$$

In fact, the same sort of argument provides a closed form formula for the generating function if we were to wait for a run of r heads, for some general $r = 1, 2, 3, \dots$. The generating function would be :

$$\psi(t) = \frac{p^r t^r (1-pt)}{1-t+qp^r t^{r+1}} \quad (2)$$

Since $E(N) = \psi'(1)$, it is a simple task to derive the neat formula

:

$$E(N) = \frac{1-p^r}{qp^r} = \frac{1}{p} + \frac{1}{p^2} + \dots + \frac{1}{p^r}. \quad (3)$$

Thus, if we were tossing a fair coin, so that p would be .5, on an average we will have to toss our coin 14 times to see a succession of three heads for the first time. But we have to toss the coin on an average only 6 times to see a succession of two heads for the first time.

Not only the expected value of N , the variance, or indeed any moment, and even the probabilities p_k can be found from the generating function $\psi(t)$. For instance, $E(N(N-1)) = \psi''(1)$, implying $Var(N) = \psi''(1) + \psi'(1) - (\psi'(1))^2$. And in general, $E(N^k) = \sum_{i=1}^k S_i^k \psi^{(i)}(1)$, with S_i^k being Stirling numbers of the second kind. A formula for the variance of N follows quite easily; indeed,

$$Var(N) = \frac{1-p^{1+2r}-qp^r(1+2r)}{q^2p^{2r}} \quad (4)$$

It is symptomatic of problems in patterns and coincidences that the variances tend to be high, as the following Table illustrates :

Table 1

r	2	3	4	5
E(N)	6	14	30	62
Var(N)	22	142	734	3390

The high variances as seen in Table 2 are consequences of a pronounced right skewed distribution for the waiting time N . Calculations using the third moment will confirm this numerically. Here is the third moment formula, somewhat complex, as might be expected.

$$E(N^3) = (6 - p^{3r} + 2p^{1+3r} - p^{2+3r} - 12p^r(1+r) + 6p^{1+r}(1+2r) + p^{2+2r}(1+3r+3r^2) - 2p^{1+2r}(4+6r+3r^2) + p^{2r}(7+9r+3r^2))/(q^3p^{3r}) \quad (5)$$

Using (5) and the variance formula in (4), the coefficient of skewness $\frac{E(N-E(N))^3}{(\text{Var}(N))^{\frac{3}{2}}}$ of the waiting time N is 2.035 if $r = 2$ and 2.010 if $r = 3$. Note the rather remarkable fact that the coefficient of skewness for an Exponential density is exactly equal to 2, very very close to the skewness coefficient for the waiting time N .

Example 3 From the generating function $\psi(t)$ in (2), the distribution of the waiting time N can be worked out too. It is interesting to look at some probabilities $P(N \leq k)$ for 2 and 3 runs for the case of a fair coin. The first row corresponds to a run of length 2 and the second row to a run of length 3.

Table 2

$P(N \leq 3)$	$P(N \leq 4)$	$P(N \leq 5)$	$P(N \leq 6)$	$P(N \leq 8)$	$P(N \leq 10)$
.375	.5	.594	.672	.785	.843
.125	.1875	.25	.3125	.418	.452

Note the interesting fact that the median waiting time for a success run of 2 is only 4 trials. Table 2 also reveals that a success run of 3 is significantly harder than a success run of 2. For example, one should observe a run of 2 successes with about an 80% probability within 8 tosses; but the probability of seeing a run of 3 successes within 8 tosses is only about half of that.

3.2 More General Patterns

It is interesting to see the generating functions for the waiting times of more general patterns on several grounds. First, in certain real applications, patterns other than runs are important. Second, generating functions of more general patterns demonstrate interesting phenomena. For example, in repeated tosses of a coin, one would always see the sequence HHT, on an average, in fewer trials than the sequence HTH; the value of p does not matter. But one would see a run of 3 heads HHH in fewer trials than the sequence HTH

only if $p > \frac{\sqrt{5}-1}{2}$, the golden number! And third, generalizations, unless too complex, are always of some theoretical interest.

Consider a sequence of iid multinomial trials with a countable number of outcomes in a single trial and general cell probabilities. Let S denote a general *pattern* $c_0c_1\dots c_r$, where each c_i is a member of our multinomial alphabet. Thus the length of the pattern S is $r + 1$. Let again N denote the waiting time till the pattern S is observed for the first time and let $p_k = P(N = k)$.

Let $\phi(t) = \sum_{n=1}^{\infty} p_n t^n$ (we do not discuss the radius of convergence here). We follow the terminology in Solov'ev(1966). However, other important references include Feller(1966), Li(1980), and Guibas and Odlyzko(1981). Call d a period of the pattern S if $c_i = c_{d+i}$ for $i = 0, 1, \dots, r - d$. For example, if in the pattern S , the first and the last letters are the same, then $d = r$ is a period of the pattern S . Let d_1, d_2, \dots, d_s be the periods of S , and let π_{d_i} denote the probability of occurrence of the last d_i letters of the pattern S in as many trials. Note that π_{d_i} is thus a function of the cell probabilities of the original multinomial experiment. Let also α_1 denote the probability of occurrence of the pattern S (the full sequence of length $r + 1$) in as many trials. Then Solov'ev(1966) presents the following formula for the generating function $\phi(t)$:

$$\phi(t) = \frac{\alpha_1 t^{r+1}}{(1-t)(1+\pi(t))+\alpha_1 t^{r+1}}, \text{ where } \pi(t) = \sum_{i=1}^s \pi_{d_i} t^{d_i} \quad (6)$$

From this formula for the generating function, the mean waiting time and the variance for a given pattern S can be worked out. Indeed,

$$E(N) = \frac{1 + \sum_{i=1}^s \pi_{d_i}}{\alpha_1}, \text{ and}$$

$$Var(N) = (E(N))^2 + \frac{2\pi'(1)}{\alpha_1} - (2r + 1)E(N). \quad (7)$$

Example 4 Consider independent tosses of a coin with probability p for heads in a single toss, and the pattern $S = \text{HTH}$. The only period of S is $d_1 = 2$. Moreover, $\pi_{d_1} = pq$. Thus, $\pi(t) = pqt^2$ and $\pi'(t) = 2pqt$. From formula (7), the mean waiting time till

HTH is observed for the first time is $\frac{1+pq}{p^2q}$ and the variance is $(\frac{1+pq}{p^2q})^2 + \frac{4pq}{p^2q} - \frac{5(1+pq)}{p^2q} = \frac{1+2p-6p^2+2p^3+3p^4-p^5}{p^4q^2}$. For $p = \frac{1}{2}$, these work out to a mean of 10 and a variance of 58.

But there is another intuitively helpful way to explain why certain patterns take longer to be observed, and others less. To do this, we need to introduce the terminology of *overlaps* as introduced in Li(1980). This is tied to the concept of periods in Solov'ev(1966). The connection of the mean waiting time to the amount of overlap in a pattern is best understood for a finite alphabet with equal cell(letter) probabilities. That is the case we will describe below.

Thus, consider a multinomial experiment with $(m + 1)$ cells described by the alphabet $\{0, 1, 2, \dots, m\}$, with equal probability $\frac{1}{m+1}$ for each cell. Let S denote a given pattern of length $k = r + 1$. We say that an overlap at level i occurs if the first i and the last i digits in the given pattern S are identical. There is an obvious connection between the concept of an overlap and the previously defined concept of a period. For instance, an overlap at level 1 occurs if and only if $d = r$ is a period for the pattern S . Let ϵ_i be the indicator of an overlap at level i . Then the mean waiting time till the pattern S is observed for the first time equals $E(N) = \sum_{i=1}^k \epsilon_i(m + 1)^i$. This formula explains why the mean waiting time is large when the pattern S has a lot of overlaps, for then many of the indicators ϵ_i are nonzero, inflating the mean waiting time. Li(1980) and Blom et.al.(1991) contain much more information on the derivation of the mean waiting time and the generating function $E(t^N)$, including some results for the case of a finite alphabet with possibly unequal cell probabilities, as well as multiple patterns (which we mention later).

Example 5 Consider all patterns of length 3 in repeated tosses of a fair coin : HHH, HHT, HTH, HTT, THH, THT, TTH, and TTT. For example, HHH has overlaps at level 1, 2, and 3; since the alphabet contains only 2 letters, here $m = 1$, and so the mean number of tosses required to observe a run of 3 heads is $2^1 + 2^2 + 2^3 = 14$, as we noted before. But consider the pattern HTH; this has

overlaps only at levels 1 and 3, and hence its mean waiting time is $2^1 + 2^3 = 10$. We see that HTH is obtained, on an average, four tosses sooner than HHH. At first glance, many find this surprising. The mean waiting times for the eight patterns of length 3 above are 14, 8, 10, 8, 8, 10, 8 and 14 respectively. Thus, of the patterns of length 3, runs are harder to obtain than anything else.

Interestingly, this is *not* the case for a coin with a general probability p of heads in a single toss.

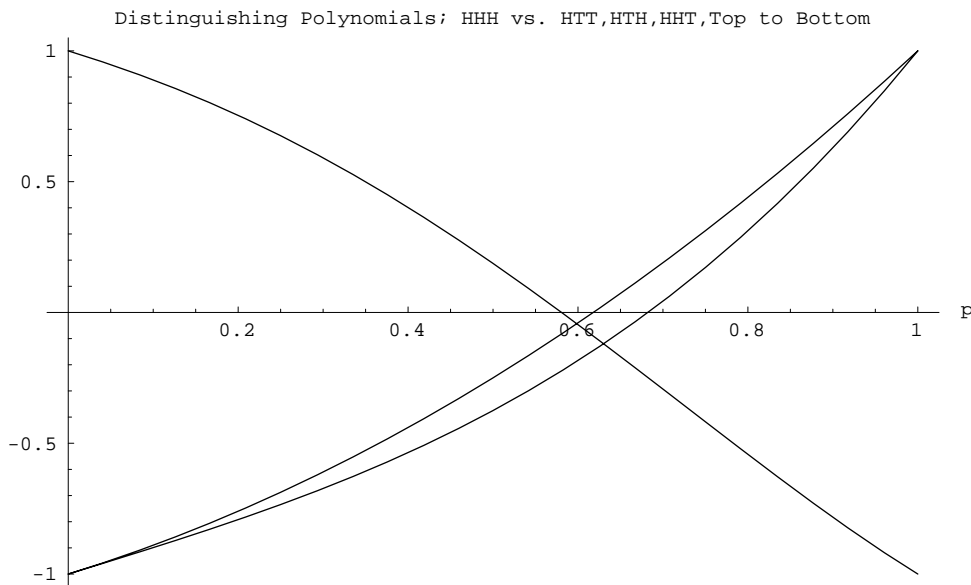
3.3 Distinguishing Polynomials

Example 6 Consider again patterns of length 3 in repeated tosses of a coin with probability p for heads in any single toss. Then the patterns HHH, HHT, HTH and HTT have mean waiting times

$$\begin{aligned} \text{HHH: } E(N) &= \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3}; \\ \text{HHT: } E(N) &= \frac{1}{p^2q}; \\ \text{HTH: } E(N) &= \frac{1+pq}{p^2q}; \\ \text{HTT: } E(N) &= \frac{1}{q^2p}. \end{aligned} \tag{8}$$

An interesting consequence of equation (8) is that HHT is always easier to obtain than HTH, regardless of the value of p ! Another very intriguing consequence is that the mean waiting time for HHH is smaller than that for HTH if and only if $p > \frac{\sqrt{5}-1}{2}$, the golden number. Thus, for adequately large p , in fact runs are easier to obtain than punctuated patterns. The occurrence of the golden number is itself a coincidence!

The values of p for which one pattern of a given length k has a smaller mean waiting time than another pattern of length k are characterized by the negative values of appropriate polynomials. We call these *Distinguishing Polynomials*. For example, take $k = 3$, and ask the question : for what values of p , is HHH easier to obtain



than HTH ? The answer is when $p^2 + p - 1 > 0$, thus leading to the golden number as the cutoff value for p . Ask the question : when is HHH easier to obtain than HTT ? The answer is when $p^4 - p^3 - p^2 - p + 1 < 0$. When is HHH easier to obtain than HHT ? The answer is when $p^3 + p - 1 > 0$. The three distinguishing polynomials $p^2 + p - 1, p^4 - p^3 - p^2 - p + 1, p^3 + p - 1$ are plotted together above. To our knowledge, no general structures have been found for these distinguishing polynomials for general values of r , the pattern length. It would be highly interesting if some general structures are found.

3.4 Dice, Monkeys and Nucleotides

Coin tosses are but one example of interest. Waiting time for patterns are of interest in many other chance experiments. Dice experiments and card games have been of interest in gambling for a very long time. They also have some purely intellectual interest. The general theory presented above can be used to study waiting times in these experiments as well. We offer a few examples.

Example 7. Consider repeated tosses of a fair die with faces marked 1,2,...,6. This is a multinomial experiment with an alphabet consisting of six letters. What are the mean waiting times to see the patterns 11, 111, 1111, 1212, 1221, and 123456 ?

It follows from the more general formulas (e.g., equation (7)) that the mean waiting times for these patterns are as follows :

Table 3

Pattern	11	111	1111	1212	1221	123456
Mean waiting time	42	258	1554	1332	1302	46656

Notice how the pattern 1221 is obtained sooner than 1212 and both are obtained substantially sooner than 1111.

Example 8 There is a common ‘joke’ that says ‘give a million monkeys a typewriter each and they will produce the writings of Shakespeare if given enough time.’ Of course, given an enormous amount of time, the monkeys will indeed do it, but the time required is fantastically long. The idea is that anything that *can* happen *will* ultimately happen. It is useful to investigate how long does it take for patterned structures to arise just by chance.

Imagine that indeed a monkey has been given a typewriter, and the monkey hits one key at random per second. To make things easier for our poor monkey, we give it a typewriter with just the 26 letters of the English alphabet. If the hits are independent, how long (in terms of time elapsed) would it take to produce the word ‘Anirban’. The only period of this word is 2 (the first two and the last two letters being the same), giving a mean value of $26^2 + 26^7 = 8031810852$ hits required on an average for our monkey to produce the word ‘Anirban’. That works out to approximately 255 years of nonstop hitting on the typewriter by our playful monkey.

Example 9 Biologists are interested for various reasons in the frequency of occurrence of specific ‘words’ along a strand of the

DNA sequence of an organism. The entire genome of an organism can be very long. It is believed that the human genome is a word consisting of about 3 billion letters, each letter being a member of the four letter alphabet $\{A,G,C,T\}$. For example, if a specific word such as GAGA occurs less or more frequently than it should under a postulated model, then that is considered to be useful information to a biologist for various reasons. It may be taken as an evidence against the postulated model, or it may be of other biological interest, for example in searching for what is called a promoter signal. See Waterman(1995) or Ewens and Grant(2001) for a wealth of information on these issues.

Consider as an example the word GAGA. Suppose we are interested in finding the mean distance along the genome for successive occurrences of this word. Then, *provided the occurrences are counted along the genome in a nonoverlapping fashion*, the mean distance can be found from formula (7), The word GAGA has only one period $d_1 = 2$, and so from formula (7), the mean renewal distance is $\frac{1+\frac{1}{16}}{\frac{1}{4^4}} = 272$. The variance follows from formula (7) too; it would be $(272)^2 + 2 \times \frac{2}{16} \times 4^4 - 7 \times 272 = 72144$, giving a standard deviation of about 270.

3.5 DNA Sequence Alignments

Identifying exactly matched or well matched sequences of DNA for two or more organisms has been at the forefront of activity in biological research for at least the last twenty years. The motivation is the following : as genetic material from an ancestor is passed on to successive generations, the sequence corresponding to this genetic material will undergo slow changes. The changes can be in the form of nucleotide deletions, insertions, or substitutions. Stretches of DNA that correspond to ‘critically important’ genes would be resistant to these mutations. Thus the rate of change depends on the functional responsibility of the specific stretch of DNA. Biologists align, either two full sequences, or parts of them, in order to understand ancestral relationship between two organisms. They also align multiple sequences to understand the relationships between

multiple organisms. These methods are useful for understanding evolutionary history. Probabilistic analysis is very helpful here, but the corresponding mathematical problems are excruciatingly difficult, mostly because some complicated dependence makes standard theory inapplicable.

As one example of the kind of analysis that is done and the mathematics involved, consider two sequences of length n each. Then we count the lengths Y_1, Y_2, \dots, Y_s of the various matched subsequences and find $Y_{(s)} = \max(Y_1, Y_2, \dots, Y_s)$. If the observed value of $Y_{(s)}$ is drastically inconsistent with what it should be were the two sequences generated independently from the alphabet $\{A, G, C, T\}$, we would conclude that there is (possibly) some relationship between the organisms. This is the main idea.

However, the problems are complicated. The smaller complication is that s , the total number of subsequences with some matching, is random. The much bigger complication is that Y_1, Y_2, \dots, Y_s are not at all independent. Extreme value theory for dependent sequences is still not well developed. Thus, computing even the mean and variance of $Y_{(s)}$ is a major challenge.

Some tentative solutions have been suggested. One line of attack is massive simulation, or using the bootstrap. A DNA strand of length n is generated repeatedly on a computer and either a P-value or the mean and variance of $Y_{(s)}$ are approximated. A more analytical approach is to give approximate formulas for the mean and variance of $Y_{(s)}$. The second approach is taken in Waterman(1995). Waterman's formulae also apply to the case of *approximate matches* wherein some mismatches in the subsequences are allowed; this is believed to be a more credible approach to studying evolutionary history rather than counting exact matches. However, use of these approximate formulas in order to plug them into an asymptotic theory result for $Y_{(s)}$ is still somewhat problematic, because the asymptotics do not seem to kick in until n , the sequence length, is very very large.

Thus, assessment of surprise by studying alignments continues

to be a very difficult probabilistic problem. Much literature already exists on this though; Waterman(1995a,b), Waterman and Vingron(1994), Griggs et.al.(1986), Karlin and Brendel(1992), Karlin and Dembo(1992), and Ewens and Grant(2001) are a few useful references.

3.6 Longest Runs, Erdős-Rényi Law and Central Limit Theorems

Suppose S is a given pattern(word) of length k . Let N_r denote the waiting time till the r -th occurrence of S and f_n the number of occurrences of S in a sequence of n iid multinomial trials. N_r and f_n are related by the obvious identity $P(f_n < r) = P(N_r > n)$. By the canonical central limit theorem for iid random variables with a finite variance, a central limit theorem for N_r follows when $r \rightarrow \infty$. From here, a central limit theorem for f_n follows under certain configurations of $r, n \rightarrow \infty$. The statements are as follows; their proofs can be seen in Feller(1966).

Theorem 1(a) If $r \rightarrow \infty$, then for any real x , $P(\frac{N_r - r\mu}{\sigma\sqrt{r}} \leq x) \rightarrow \Phi(x)$, where μ, σ are the mean and the standard deviation of the waiting time till the first occurrence of S (formula (7)) and $\Phi(\cdot)$ is the standard normal CDF.

(b) If $r, n \rightarrow \infty$ in such a way that $\frac{n - r\mu}{\sigma\sqrt{r}} \rightarrow x$, then $P(f_n \geq r) \rightarrow \Phi(x)$, or equivalently, $P(\frac{f_n - \frac{n}{\mu}}{\sqrt{n\frac{\sigma^2}{\mu^3}}} \leq x) \rightarrow \Phi(x)$.

Loosely speaking, it is a consequence of the renewal theorem that to the first order, f_n is of the order of $\frac{n}{\mu}$. For fixed r and n , the distribution of f_n can be found, probably with heavy computation, from the generating function for N , the waiting time for the first occurrence, as given in formula (6), and on using the aforementioned relation $P(f_n < r) = P(N_r > n)$. Unless r, n are adequately small, a direct computation such as this would be at least time consuming and probably frustrating. Thus, the asymptotics are evidently useful.

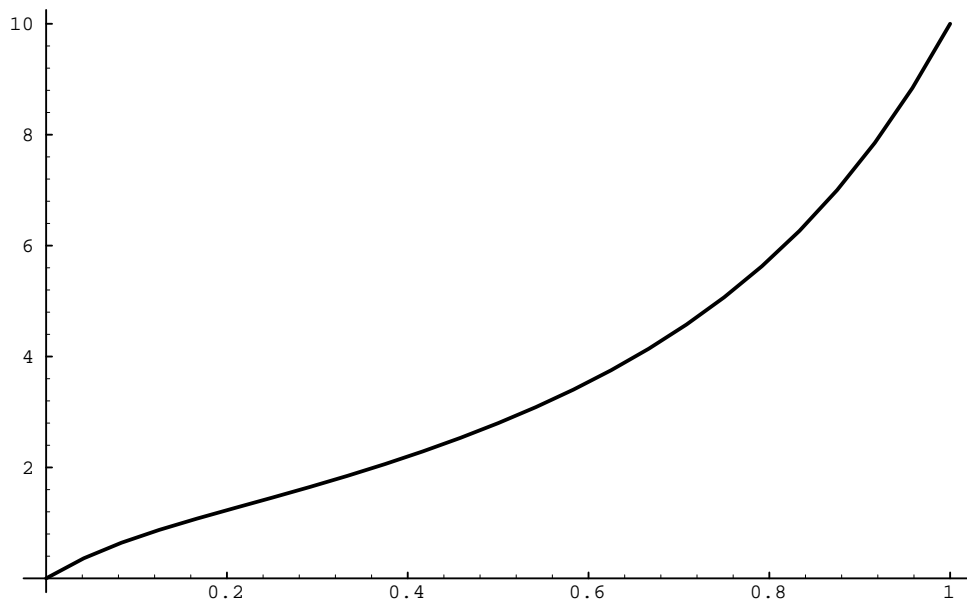
Another related quantity of great theoretical and practical interest is the longest run of one letter in n iid multinomial trials. A simple example is the length of the longest head run in n independent tosses of a coin. The longest run tends to be quite a bit higher than (most) people naively expect. In experiments where people are asked to write outcomes of coin tosses as part of a mind experiment, most people never write a run longer than 3. So the classic Erdős-Rényi result that pins down the rate of growth of the longest run is extremely interesting.

Theorem 2 Consider iid multinomial trials with $p = P(l)$, l being a specific letter in the multinomial alphabet. Let L_n denote the length of the longest run of the letter l in n trials. Then $\frac{L_n}{\log_{\frac{1}{p}} n} \xrightarrow{a.s.} 1$.

If for fixed k , $p_n = P(L_n \geq k)$ and $u_k(t) = \sum_{n=k}^{\infty} p_n t^n$, then a recursion relation on the sequence $\{p_n\}$ shows that $u_k(t) = \frac{p^k t^k (1-pt)}{(1-t)(1-t+qp^k t^{k+1})}$.

It follows that $p_n = \frac{u_k^{(n)}(0)}{n!}$, and $E(L_n) = \frac{1}{n!} \sum_{k=1}^n u_k^{(n)}(0)$. For relatively small n , one can compute the values of p_n and $E(L_n)$ from these relations. There is no good way to compute them exactly for large n . However, in the case $p = \frac{1}{2}$, the approximation $p_n \approx e^{-n2^{-k-1}}$ is often accurate and from here one can conclude that $E(L_n) = \log_2 n +$ a small periodic component. See, for example, Odlyzko(1995).

The next figure plots $E(L_n)$ as a function of p when $n = 10$. The expectation has a sharp increase around $p = .6$. For the case of $p = .5$, the expected value of L_n is given in the following Table for a few small values of n .



Example 10 Table 4

n	6	10	15
p			
.25	1.13	1.46	1.71
.4	1.71	2.20	2.60
.5	2.16	2.80	3.34
.6	2.67	3.54	4.27
.75	3.63	5.07	6.34

The values in Table 4 indicate that if the coin was a bit biased towards heads, some clumping phenomena would be observed. Even with six tosses, a head run of three would not be unusual, while one would expect to see only about four heads in total.

3.7 Multiple Patterns

A typical question involving multiple patterns is how many tosses of a coin are required before a head run of r or a tail run of s is observed. Such questions are of interest in statistical process control. Multiple patterns are also of interest to biologists, who look for frequency of occurrence of several words in a strand of DNA, a word being a sequence of nucleotides.

Using martingale methods, and the optional sampling theorem in particular, Li(1980) gave a very ingenious solution to the problem of calculating the mean waiting time in a sequence of iid multinomial trials until one of several patterns is observed for the first time. The special case of a head run or a tail run in iid coin tosses had been known earlier, and Feller(1966) is a standard reference.

Let S_1, S_2, \dots, S_p be p patterns (words) constructed from a multinomial alphabet $\{0, 1, \dots, m\}$, the digits $0, 1, \dots, m$ having equal probability $\frac{1}{m+1}$. The patterns S_i may have different lengths, but here we consider only the case where the patterns all have the same length, say k . Let N be the number of trials required for one of the patterns S_1, S_2, \dots, S_p to be observed for the first time. Then, using concepts of *overlaps* analogous to the one-pattern case, Li(1980) characterizes $E(N)$ in the following manner.

Theorem 3 Let $\epsilon_n(i, j)$ be the indicator of the last n digits in S_i being identical to the first n digits in S_j . Let $\mu_{ij} = \sum_{n=1}^k \epsilon_n(i, j)(m+1)^n$. Then $E(N)$ and the p probabilities $\pi_i = \mathbf{P}(\text{The experiment terminates with } S_i \text{ being the pattern first observed})$ are together found by solving the system of linear equations :

$$\begin{aligned} \sum_{i=1}^p \epsilon_{ij} \pi_i &= E(N) \quad (\text{for every } j) \\ \sum_{i=1}^p \pi_i &= 1. \end{aligned} \quad (9)$$

Example 11 Consider the classic example of waiting for a head run of length r or a tail run of length s in iid coin tosses. Although only the fair coin case would follow from Theorem 3 above, the general case has been known for a long time. The mean waiting

time is given by

$$E(N) = \frac{(1-p^r)(1-q^s)}{qp^r + pq^s - p^r q^s}. \quad (10)$$

For example, the mean number of tosses needed to see either a run of three heads or a run of three tails in iid tosses of a fair coin is *only* 7. The mean number of tosses needed to see either a run of four heads or a run of four tails is 15.

4 Birthday and Strong Birthday Problems

The classical birthday problem that asks what is the probability of finding at least one similar pair in a group of n individuals, where a similar pair is a pair with the same birthday, was initiated by von Mises in 1932. The strong birthday problem is a word coined by us, and asks what is the probability that everyone 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 more than 50% is $n = 23$. In the strong birthday problem, the smallest n for which the probability that everyone has a shared birthday is more than 50% 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.

4.1 The Canonical Birthday Problem

For a group of n individuals, let I_{ij} be the indicator of the event that individuals i, j have the same birthday. Then, the number of similar pairs $W = \sum_{i < j} I_{ij}$, a sum of Bernoullis. But 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 :

Table 5

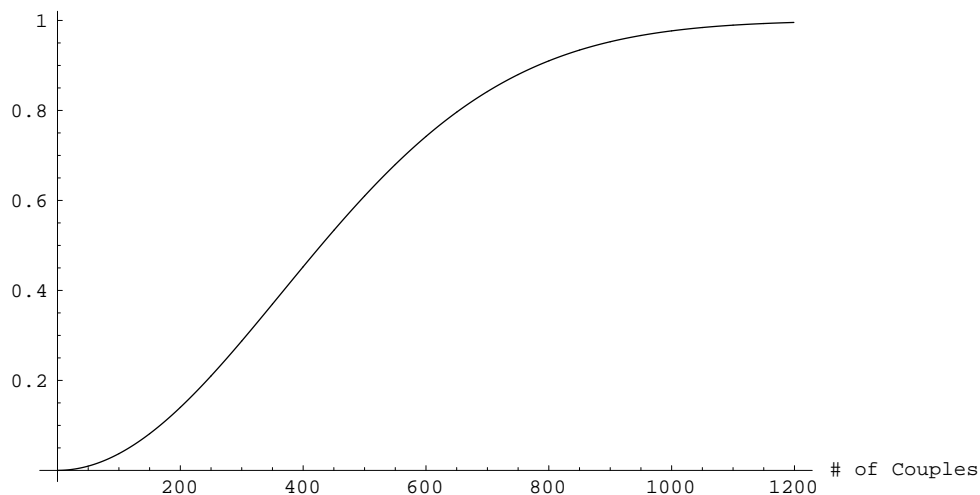
n	2	3	4	5	10	20	23	30	50
p_n	.003	.008	.016	.027	.117	.411	.507	.706	.970

Thus, in a gathering of 50 people, it is virtually certain 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, well enough that the Poisson approximation to the probability of at least one similar pair gives almost exactly the $n = 23$ value as the n required for the probability to exceed .5. The following result can be seen in various places, including Arratia et. al.(1989,1990), Stein (1986), Barbour et. al. (1992), and Diaconis and Holmes (2002).

Theorem 4 If $m, n \rightarrow \infty$ in such a way that $\frac{n(n-1)}{m} \rightarrow 2\lambda$, then $W \xrightarrow{\mathcal{L}} Poi(\lambda)$.

If m, n are 365 and 23 respectively, then taking 2λ as $\frac{22 \times 23}{365} = 1.3863$, and using the Poisson approximation, $P(W \geq 1) = 1 - e^{-.69315} = .500001$, an accurate approximation to the true value of .507 (Table 5).

A slightly different question is the following : if we interview people, one by one, until we have found a similar pair, how many would need to be interviewed ? If we denote the number we will need to interview 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.



4.2 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 $n = 431$. A plot of the probability of finding at least one pair of matched couples is given above.

4.3 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

$$p(m, n, k) = \frac{(m-nk-1)!}{(m-n(k+1))!m^{n-1}}. \quad (11)$$

Example 12 Using this formula, 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).

4.4 Similar Triplets and p of a Kind

The 23 answer 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 $P(W \geq 1) \geq .5$. 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 $P(W \geq 1)$:

$$P(W \geq 1) = 1 - \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{m!n!}{i!(n-2i)!(m-n+i)!2^i m^n}. \quad (12).$$

Example 13 If $m = 365$, $P(W \geq 1)$ is .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

$$E(N_p) \sim (p!)^{\frac{1}{p}} \Gamma(1 + \frac{1}{p}) m^{1 - \frac{1}{p}}, \quad (13)$$

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

4.5 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, \dots, p_m) of the m birthdays is assumed unknown and a prior distribution on it, or equivalently, on the $(m - 1)$ -dimensional simplex Δ_m is assumed. The questions of main 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 $n \rightarrow \infty$.

If the vector of cell probabilities has an exchangeable *Dirichlet* $(\alpha, \alpha, \dots, \alpha)$ prior, then $P(W \geq 1) = 1 - \prod_{i=1}^{n-1} \frac{\alpha(m-i)}{m\alpha+i}$. 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 $P(W \geq 1)$ to be $\geq .5$ 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	hspace.1in21	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 5 Suppose $m, n \rightarrow \infty$ in such a way that $\frac{n(n-1)}{m} \rightarrow 2\lambda$. Then, under the exchangeable *Dirichlet*($\alpha, \alpha, \dots, \alpha$) prior, $W \xrightarrow{\mathcal{L}} Poi(\frac{\alpha+1}{\alpha}\lambda)$.

Thus, under the uniform prior, one would expect about twice as many similar pairs as in the classical case. This is very interesting.

4.6 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 ? Colloquially, we may state this as what is the probability that none in a group of n people is *unique* ? If we let the number of unique people be N , then the problem asks what is $P(N = 0)$? More generally, one can ask what is $P(N = k)$? 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 $S_i = \frac{m!n!(m-i)^{n-i}}{i!(m-i)!(n-i)!m^n}$,

$$P(N = k) = \sum_{i=k}^n (-1)^{i-k} \frac{i!}{k!(i-k)!} S_i. \quad (14).$$

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.

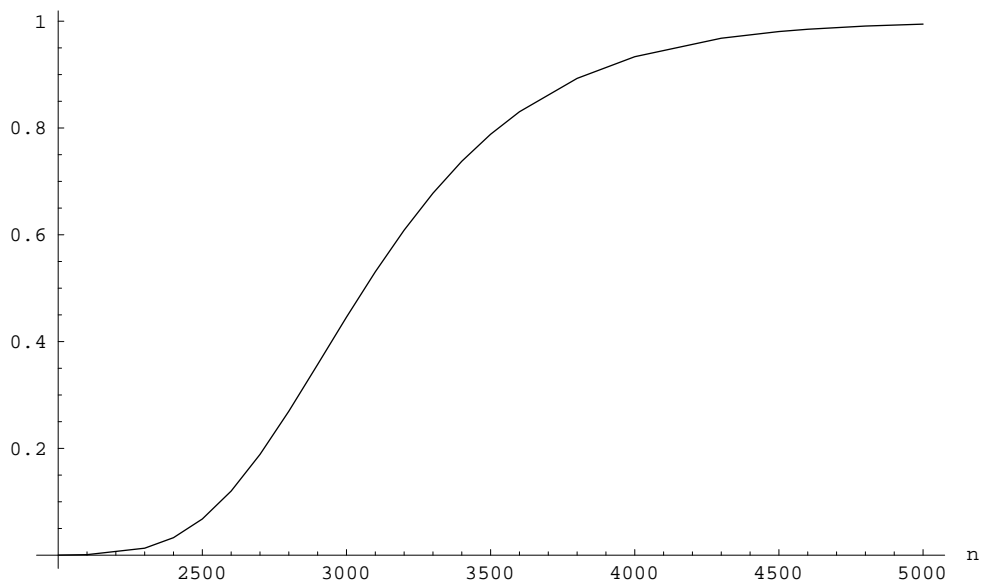
Example 14 Table 6

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

Thus, in a group of 3064 people, the odds are better than half that everyone has a shared birthday. A plot of p_n is given below.

An accurate iterative approximation of the smallest n required to make $p_n = p$ is $\frac{n_1}{m} = \log\left(\frac{m}{\log \frac{1}{p}}\right)$, $\frac{n_i}{m} = \frac{n_1}{m} + \log\left(\frac{n_{i-1}}{m}\right)$. For $p = .5$, the iterations produce the values 2287,2957,3051,3062,3064. Thus, five iterations give the correct value of n .

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 6 Suppose $\frac{n}{m} = \log n + c + o(1)$. Then, under the equal probability and independence (of the people) assumptions, $N \xrightarrow{\mathcal{L}} Poi(e^{-c})$.

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

4.7 Bayesian Versions

For general arbitrary birthday probabilities p_1, p_2, \dots, p_m , the distributional properties of N can get very complicated and even strange. Of course, the mean and the variance are easy :

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

$$Var(N) = n(n-1) \sum_{i \neq j} p_i p_j (1 - p_i - p_j)^{n-2} + E(N) - (E(N))^2. \quad (15)$$

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

$$E_\alpha(N) = \frac{m\alpha\Gamma(m\alpha)\Gamma(\alpha(m-1)+n-1)}{\Gamma((m-1)\alpha)\Gamma(m\alpha+n)} \quad (16)$$

As regards the marginal probability $P(N = 0)$, for a fixed vector of cell probabilities p_1, p_2, \dots, p_m , it equals $1 - \sum_{j=1}^{\min(m,n)} (-1)^{j-1} \frac{n!}{j!(n-j)!}$

$\times \sum_{i_1 \neq i_2 \neq \dots \neq i_j} p_{i_1} p_{i_2} \dots p_{i_j} (1 - p_{i_1} - p_{i_2} - \dots - p_{i_j})^{n-j}$. This can be integrated with respect to a $Dirichlet(\alpha, \alpha, \dots, \alpha)$ density on using the Liouville integral formula $\int_{\Delta_s} p_1^{a_1} p_2^{a_2} \dots p_s^{a_s} f(p_1 + p_2 + \dots + p_s) dp_1 \dots dp_s = \frac{\Gamma(1+a_1)\dots\Gamma(1+a_s)}{\Gamma(s+a_1+\dots+a_s)} \int_0^1 x^{s+a_1+\dots+a_s-1} f(x) dx$. The resulting expressions are still closed form, but messy.

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

$$P_u(N = 0) = 1 - \frac{(m-1)!m!n!}{(m+n-1)!} \sum_{j=1}^{m-1} (-1)^{j-1} \frac{(m+n-2j-1)!}{j!(n-j)!(m-j-1)!(m-j)!}. \quad (17)$$

However, for large values of m, n , due to the alternating nature of the sum in formula (17), the computation appears to turn unreliable. Thus, asymptotics would be useful again. From formula (17), one can show that if $n \sim \theta m^2$, then $P(N = 0) \rightarrow e^{-\frac{1}{\theta}}$. 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.

4.8 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 15 Using a set of 14 physical features, such as sex, height, built, facial shape, nasal elevation, size of pupils, eyebrow thickness, size of head, etc., with 2 to 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.

5 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 $1, 2, \dots, 52$ 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 quite a bit of the elegant combinatorics of *random permutations*.

5.1 Unrestricted Matching Problem

The problem can be formulated as follows. Let π denote a random permutation of the n numbers $1, 2, \dots, n$; i.e., the probability that π is any of the $n!$ permutations of $\{1, 2, \dots, n\}$ 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 $0, 1, \dots, n$.

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 $n \rightarrow \infty$ to $e^{-1} \approx .36788$. 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 16 Consider the distribution of Z when $n = 10$. By using the formula given above, the distribution is as follows.

Table 7

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

The rest of the probabilities are too small. The most striking aspect of Table 7 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 . It is interesting that these empirical phenomena are in fact analytically provable. See DasGupta(1999) and also consult Diaconis and Holmes(2002).

5.2 Error of Poisson Approximation

Theorem 7 (a) Let $d_{TV} = \frac{1}{2} \sum_{k=0}^{\infty} |P(Z = k) - P(Poisson(1) = k)|$. Then d_{TV} admits the integral representation

$$d_{TV} = \frac{1}{n!} \int_0^1 e^{-t} t^n dt + \frac{1}{(n-1)!} \int_0^1 (2-t)^{n-1} t^{-n-1} \gamma(n+1, t) dt, \quad (18),$$

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

(b) $d_{TV} \leq \frac{2^n}{(n+1)!}$ for every $n \geq 2$.

The integral representation in (a) implies the error bound in (b). The error bound explains why the Poisson approximation in Table 7 is so sharp. Although much easier proofs can be given, the error bound also implies that $Z \xrightarrow{\mathcal{L}} Poisson(1)$ as $n \rightarrow \infty$. 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 8 (a) For n even, $P(Z = k) < (>)P(Poisson(1) = k)$ according as $k < n$ is odd or even; the opposite inequalities hold for n odd.

(b) The inequalities of part (a) hold when $(Z = k), (Poisson(1) = k)$ are replaced by $(Z \leq k), (Poisson(1) \leq k)$.

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.

5.3 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, a very neat result holds.

Theorem 9 Suppose the size N of the deck is distributed as $Geometric(p)$, with mass function $p(1-p)^n, n \geq 0$ (thus, an empty

deck is allowed). Then the (marginal) distribution of Z is *exactly* a *Poisson*, with mean $1 - p$.

If p is parametrized by m , with $p_m \rightarrow 0$, then, still, $Z \xrightarrow{\mathcal{L}} \text{Poisson}(1)$.

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 9. 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.

Example 17	Table 8 : P(No Matches)	
Expected Deck Size	Uniform Deck	Geometric Deck
5	.315	.4346
25	.357	.382
50	.3626	.375

5.4 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 $\{1, 2, \dots, n\}$, what is the probability that $|\pi(i) - i| \leq 1$ for every $i = 1, 2, \dots, n$? More generally, one can ask what is $P(\max_{1 \leq i \leq n} |\pi(i) - i| \leq r)$, where $r \geq 0$ is a fixed integer ?

Example 18 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 n th 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 $n = 2, 3, \dots, 9$ respectively. Thus, calling the numbers within an error of 2 should already be of considerable surprise if n is 8.

See Tomescu(1985) for these results.

5.5 Longest Increasing Subsequence

Consider a random permutation of $\{1, 2, \dots, n\}$. 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 (Erdős and Szekeres(1935)); actually Erdős 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 $\{1, 2, \dots, n\}$. Then, $\frac{I_n}{\sqrt{n}}$ converges in probability to 2, and furthermore, $\frac{E(I_n)}{\sqrt{n}} \rightarrow 2$. In fact, even second order asymptotics for $E(I_n)$ are known; settling a longstanding conjecture founded on Monte Carlo and other evidence, Baik, Deift and Johansson(1999) established the result $\frac{E(I_n) - 2\sqrt{n}}{n^{\frac{1}{6}}} \rightarrow c_0$, where c_0 is the mean of the *Tracy-Widom distribution* on the reals. An approximate numerical value for c_0 is -1.7711 (there seems to be a typographical error in Aldous and Diaconis(1999)). The CDF of the Tracy-Widom distribution does not have a closed form formula, but numerical evaluation is possible, by numerical solution of a corresponding differential equation. In fact, one has the

remarkable result that $\frac{(I_n - 2\sqrt{n})}{n^{\frac{1}{6}}} \xrightarrow{L} L$, L having the Tracy-Widom distribution. See Tracy and Widom(1994), Baik, Deift and Johansson(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 tableau*. Evolution of these results can be seen in Hammerley(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. These results provide precise mathematical foundations for deciding what would cause a surprise in seeing a long increasing subsequence of a random permutation.

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.

5.6 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.

A permutation π of $\{1, 2, \dots, n\}$ 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). 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 $\frac{n+1}{2}$.

A permutation π is called an up-down permutation if $\pi(1) < \pi(2) > \pi(3) < \pi(4) > \dots$; 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 $\{1, 2, \dots, n\}$, 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 $\text{sect} + \text{tant}$; see Andre(1879) or Tomescu(1985). Thus, $u_n = \frac{d^n}{dt^n}(\text{sect} + \text{tant})|_{t=0}$. For example, $u_5 = 16$, and $u_{10} = 50521$. The probability of seeing an up-down permutation is listed in the following Table for some selected values of n ; it would be an extreme surprise to observe one if n was 15 or so.

Example 19

Table 9

n	$P(\text{An Up-down Permutation})$
2	.5
3	.333
4	.208
5	.133
7	.054
10	.014
15	.001
20	.00015

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

References

- Abramson, M. and Moser, W. (1970). More birthday surprises, *Amer. Math. Monthly*, 77, 856-858.
- Aldous, D. and 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. and Gordon, L. (1989). Two moments suffice for the Poisson approximation: The Chen-Stein method, *Ann. Prob.*, 17, 1, 9-25.
- Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen-Stein method, *Statist. Sc.*, 5, 403-434.
- Baer, R. M. and Brock, P. (1968). Natural sorting over permutation spaces, *Math. Comp.*, 22, 385-410.
- Baik, J., Deift, P. and Johansson, K. (1999). On the distribution of the length of the longest increasing subsequence of random permutations, *J. Amer. Math. Soc.*, 12, 1119-1178.
- Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, Oxford University Press, New York.
- Blom, G., Holst, L. and Sandell, D. (1991). *Problems and Snapshots from the World of Probability*, Springer-Verlag, New York.
- Cole, K. C. (2003). *Mind Over Matter*, Harcourt, Inc., San Diego.
- DasGupta, A. (1999). The matching problem with random decks and the Poisson approximation, Tentatively accepted, *Sankhyá*, Ser. A.
- DasGupta, A. (2001). Strong birthday problems and look-alikes. Preprint, Purdue University.
- Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences, *JASA*, 84, 408, 853-861.

Diaconis,P. and Holmes,S.(2002). A Bayesian peek at Feller Volume I, Sankhy a, Ser. A, Special Issue in Memory of D. Basu, Ed. Anirban DasGupta, 64,3(ii) , 820-841.

Erdős,P. and Szekeres,G.(1935). A combinatorial theorem in geometry, Comp. Math, 2, 463-470.

Ewens,W.J. and Grant, G.R.(2001)*Statistical Methods in Bioinformatics*, Springer, New York.

Feller, W.(1966). *An Introduction to Probability Theory and Its Applications, I*, John Wiley, New York.

Griggs,J.R.,Hanlon,P.J. and Waterman,M.S.(1986). Sequence alignments with matched sections, SIAM J. Alg.Discrete Meth., 7, 604-608.

Guibas, L.J. and Odlyzko, A.M.(1981). String overlaps, pattern matchings and nontransitive games, J.Comb.Theory, A, 30, 183-208.

Hammersley,J.M.(1972). A few seedlings of research, Proc. Sixth Berkeley Symp. Math. Statist, and Prob., I, 345-394, Univ. California Press.

Karlin,S. and Brendel,V.(1992). Chance and statistical significance in protein and DNA sequence analysis, Science, 257, 39-49.

Karlin,S. and Dembo,A.(1992). Limit distributions of maximal segmental scores among Markov-dependent partial sums, Adv. Appl. Prob., 24, 113-140.

Klamkin,M.S. and Newman,D.J.(1967). Extensions of the birthday surprise, J. Comb. Theory,3,279-282.

Kolchin,V.F.,Sevast'yanov,B.A., and Chistyakov,V.P.(1978). *Random Al locations*, John Wiley, New York.

Lawler,G.(1996).*Intersections of Random Walks*, Birkhauser,

Berlin

Le Cam, L.(1960).An approximation theorem for the Poisson-Binomial distribution, *Pacific J. Math*, 10, 1181-1197.

Li, S.R.(1980). A martingale approach to the study of occurrence of sequence patterns in repeated experiments, *Ann. Prob.*, 8, 1171-1176.

Logan,B.F. and Shepp,L.(1977). A variational problem for random Young tableaux ,*Adv. in Math*,26, 206-222.

Odlyzko,A.M.(1995). Asymptotic enumeration methods, In *Handbook of Combinatorics*, II, Eds. R.L. Graham, M. Grötschel and L. Lovász, MIT Press, North-Holland.

Solov'ev,A.D.(1966). A combinatorial identity and its application to the problem on the first occurrence of a rare event, *Teor. Veroyatnost.i Primenen*,11, 313-320.

Stein,C.(1986). *Approximate Computation of Expectations*, *IMS Monograph Series*, Hayward, California.

Steele,J.M.(1995). Variations on the long increasing subsequence theme of Erdős and Szekeres, In *Discr. Prob. and Algorithms*, Eds. D.Aldous, P.Diaconis and J.M. Steele, Springer-Verlag, 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*, *John Wiley*, New York.

Tracy,C.A. and 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, I n Modern Mathematics for Engineers, Ed. E.F.Beckenbach, McGraw Hill, New York.

Versik,A.M. and Kerov,S.(1977).Asymptotics of the Plancherel measure of the sy mmetric group and the limiting form of Young tables,Soviet Math Dokl.,18, 527- 531.

Waterman,M.S. and Vingron,M.(1994). Sequence comparison significance and Poiss on approximation, Statist. Sc., 9,3, 367-381.

Waterman,M.(1995a). *Introduction to Computational Biology*, Chapman and Hall, New York.

Waterman,M.(1995b). Applications of Combinatorics to Molecular Biology, In Han dbook of Combinatorics, II, Eds. R.L. Graham, M. Grötschel and L. Lová sz, MIT Press, North-Holland.