J. R. Statist. Soc. A (1997)
160, *Part* 2, *pp*. 187–206

The Statistics of the National Lottery

By JOHN HAIGH†

University of Sussex, Brighton, UK

[*Read before* The Royal Statistical Society *at an Ordinary Discussion Meeting on the National Lottery on Wednesday, November 13th, 1996, the President,* Professor A. F. M. Smith, *in the Chair*]

SUMMARY

Some methods to test the randomness of the numbers drawn in the UK National Lottery are described. The data for the first 96 draws are consistent with the numbers being drawn at random, both in terms of the individual frequencies and of the waiting times for successive appearances. The number combinations chosen by gamblers are shown to be far from random, as a whole: data from the UK and other countries demonstrate that certain combinations are much more popular than average, and thus the skill to identify unpopular combinations can increase the mean return. Families of models of gambler choice are described, but, despite some encouraging indications, they all have significant deficiencies.

Keywords: GOODNESS OF FIT; LOTTERIES; MODELS OF GAMBLER CHOICE

1. INTRODUCTION

The main National Lottery game in the UK has the same structure as 'lotto' games conducted elsewhere, for which data generated over many years can be obtained. A gambler pays £1 for the right to select six different numbers from the list {1, 2, 3, . . ., 49} and will win a prize if at least three of these choices are in common with the numbers on six similarly numbered balls, drawn 'at random' from a set of 49 identically composed balls. The order of the numbers is immaterial. A seventh 'bonus' ball is drawn but is relevant only when a gambler matches exactly five of the numbers on the main balls drawn. Fuller details are given in Moore (1997).

The data that are available for the UK National Lottery have been the numbers drawn, the numbers of tickets sold and the numbers of prize-winners (plus the sizes of the prizes) in the first 96 draws from November 19th, 1994, to September 14th, 1996. Up-to-date information can be found on the Internet site

`http://www.connect.org.uk/lottery`

shortly after each draw. For the Canadian lotto 6/49, which is conducted on almost identical lines, we have the corresponding data for all 1211 draws made from its inception to the end of August 1995, except for the number of tickets sold and the number of 'match 3' prizes. But we can give reasonable estimates of the number of tickets sold from knowledge of the prize structure, the numbers of prize-winners and the sizes of the prizes. Riedwyl (1990) has published the winning combinations for 20 years of the Swiss lotteries and extensive data on gambler choice in one particular draw.

†*Address for correspondence*: Centre for Statistics and Stochastic Modelling, School of Mathematical Sciences, University of Sussex, Falmer, Brighton, BN1 9QH, UK.
E-mail: J.Haigh@sussex.ac.uk

 0964–1998/97/160187

2. ARE THE NUMBERS DRAWN ‘RANDOM’?

The gambling public needs to be assured that the draw is conducted fairly, and that all the number combinations have an equal chance of being selected. Whatever physical checks are made on the process, and whatever steps are taken to ensure the integrity of the draw, the best assurance comes from statistical considerations: do the numbers drawn pass reasonable tests of ‘randomness’? Bellhouse (1982a, b) reported that the different Canadian lottery corporations varied widely in their quality control procedures, with some lottery data suggesting that true randomness was not being achieved. Camelot, the UK lottery franchisees, commission an independent body, BSI Test, to carry out tests both on the numbers drawn live and others using the lottery machines under controlled conditions, but they do not disclose the specific statistical outcomes. These tests are outlined in Appendix A.

Joe (1993) examined the question of testing randomness for an m/M lotto, i.e. when m numbers are to be selected at random from the integers $\{1, 2, \ldots, M\}$. Let

$$M^* = \binom{M}{m}$$

denote the total number of possible combinations, and let $B_1(d) < \ldots < B_m(d)$ denote the m numbers selected in draw d, for $d = 1, 2, \ldots, D$. The null hypothesis H_0 is that all M^* combinations are equally likely, and the draws are independent. Joe derived tests of uniformity for the distributions of single numbers, pairs and triples, and of independence between draws. He pointed out (as did Stern and Cover (1989)) that the usual goodness-of-fit statistic $\Sigma\,(O_k - E_k)^2/E_k$ needs modification if it is to be used with the χ^2-distribution; in the univariate case this is achieved by a scale factor (see test A below), but the other tests need additional terms to account for the overlaps between distinct k-tuples. The test statistics for m/M lottos are given explicitly in Joe's work, and their distributions are shown to be asymptotically χ^2. For a 6/49 lottery, this test statistic for a uniform distribution of *pairs* $\alpha = (i, j)$, where $i < j$, within the same draw is

$$W = \left\{ 90.59 \sum_{\alpha} (O_\alpha - E)^2 - 1.632 \sum_{m(\alpha,\beta)=1} (O_\alpha - E)(O_\beta - E) \right\} \Big/ D$$

as a χ^2_{1175}-distribution. Here $E = 5D/392$, and the second sum is over all (α, β) which have exactly one member in common. Similarly, to test independence between pairs of successive draws, write $\alpha = (i, j)$ with i in draw d and j in draw $d+1$. The test statistic is

$$W = \left\{ 80.22 \sum_{\alpha} (O_\alpha - E)^2 - 1.443 \sum_{u(\alpha,\beta)=1} (O_\alpha - E)(O_\beta - E) \right\} \Big/ (D-1)$$

as a χ^2_{2400}-distribution. Here $E = 36(D-1)/2401$, and the second sum is over (α, β) in which either the first or the second (but not both) members of α and β are the same.

It will be several years before D is sufficiently large for a meaningful test from this family to be made on the data from the UK National Lottery, other than on individual frequencies. However, Joe's (1993) tests, applied to several thousand draws, seem appropriate for the verification that the lottery equipment behaves as intended. They would also be useful to test the lucky dip facility, available from draw 71 on March 17th, 1996, whereby gamblers can allow Camelot's computers to choose their selections supposedly at random.

Johnson and Klotz (1993) took a different approach for data on a similar American lottery. They supposed that ball i had probability p_i of being selected first, and the probabilities on the second and subsequent selections were simply rescaled to take account of the balls already drawn. Their data consisted of the numbers drawn, in the order that they were drawn, over 200 draws. They tested the null hypothesis of uniformity, using the usual log-likelihood statistic, noting that this approach has the twin advantages of taking account of the information in the order of selection and of giving a clear alternative model if the null hypothesis is rejected. Their data gave mild evidence *against* uniformity (p-value 0.084), which they linked with the observation that the balls enter the mixing machine in the same order each draw (as happens in the UK).

Morgan (1984) noted that the criterion of 'fitness for purpose' may suggest certain specific tests on sets of allegedly random numbers, in addition to obvious tests such as equality of frequency and lack of pairwise dependence, but that there can be no expectation of full agreement on a unique, universally acceptable set of tests. For m/M lottos, there are several alternative hypotheses to the null hypothesis of equality of frequency that we might wish to have high power to detect:

(a) HA1—one number has probability $m/M+\epsilon$ of selection; the other $M-1$ have probability $m/M-\epsilon/(M-1)$, for some $\epsilon \neq 0$;
(b) HA2—half the numbers have probability $m/M+\delta$; the other half have probability $m/M-\delta$, for some $\delta \neq 0$;
(c) HA3 or HA4—for each $i = 1, 2, \ldots, M$, given that i has been selected, then either HA1 or HA2, with (m, M) replaced by $(m-1, M-1)$.

If it proves reasonable to assume equality of frequency, then tests of independence between draws can be based on the distribution of the lengths of intervals between occurrences of each and every number i. Note that, since m numbers are selected each draw, the whole collection of interval lengths are not independent. (After one draw, there must be m values, all equal to 1.)

We offer several possible tests and report the results of applying them to the six main numbers drawn in the UK lottery in Table 1.

2.1. *Test A: Equality of Marginal Frequencies*

In D draws, suppose that ball k is drawn X_k times; under hypothesis H_0, $E(X_k) = mD/M$, and the usual goodness-of-fit statistic must be modified, as noted earlier, to adjust for the sampling of m balls at a time without replacement. This statistic reduces to

$$\frac{M(M-1)\left(\sum X_k^2 - m^2D^2/M\right)}{(M-m)Dm} = W \qquad \text{(say)} \qquad (1)$$

to be compared with a χ^2_{M-1}-distribution (when D is sufficiently large, of course). It is easy to calculate that

TABLE 1

Data relating to tests A–D, applied to the six main numbers drawn in the first 96 draws of the UK National Lottery

Test A: individual frequencies of 1 → 49†

8	14	9	14	17	11	12	8	9	9	15	12	10	15	12	13	15	13	7	7	10	13	10	7	15
14	9	16	11	13	12	13	13	11	10	8	8	14	4	12	13	15	11	21	14	12	12	15	10	

Test B‡

Gap size	1	2	3	4	5	6	7	8	9	10	11	12	13
Observed	71	58	65	45	48	28	41	27	27	21	16	19	20
Expected	75.78	65.88	57.23	49.7	43.19	37.54	32.64	28.35	24.61	21.38	18.58	16.1	13.96
Gap size	14	15	16	17	18	19	20	21–22	23–24	25–27	28–31	⩾ 32	
Observed	16	13	9	8	7	7	0	7	8	7	4	4	
Expected	12.12	10.52	9.13	7.9	6.87	5.93	5.16	8.37	6.27	6.6	5.38	6.81	

Test C

$U = 152.4$ $Z = 0.716$ $V = 1189.3$ $(D-1)V/\sigma^2 = 105.10$ (95 degrees of freedom)

Test D§

No. of even numbers	0	1	2	3	4	5	6
Observed	0	2	25	38	21	10	0
Expected	1.22	8.75	23.97	31.96	21.88	7.29	0.92

†The χ^2_{48} goodness-of-fit statistic for equal mean frequencies has the value 43.97.
‡χ^2-statistic (24 degrees of freedom) 21.07.
§Goodness-of-fit statistic 7.98 on 4 degrees of freedom.

$$E(W|\text{HA1}) = M - 1 + \frac{\epsilon^2 M^2 (D-1)}{m(M-m)},$$

and that

$$E(W|\text{HA2}) = M - 1 + \frac{\delta^2 M(M-1)(D - M/m)}{M-m}.$$

Hence, for example, a conventional 5% significance test applied to data from a 6/49 lotto would need about $1.8/\epsilon^2$ draws to have a 50% chance of detecting HA1.

2.2. *Test B: Independence between Draws*

For any fixed number i, let W_1 denote the number of draws until i first appears. Similarly, let W_2, W_3, . . . be the numbers of draws between later successive appearances of i. Under hypothesis H_0 any W_r has the geometric distribution,

$$P(W_r = k) = \left(\frac{M-m}{M}\right)^{k-1} \frac{m}{M},$$

with the values of W_r relating to any fixed i being independent (but not, as noted, the whole set of Ws across the M numbers). For each i, we have observations on $W_1(i)$, $W_2(i)$, . . ., $W_{x(i)}(i)$, and a 'censored' value $W_0(i)$ (say) that denotes 1 plus the number of draws since the last appearance of i. The identities

$$\sum_{r=0}^{x(i)} W_r(i) = D \qquad \text{(for each } i \text{ in turn)}$$

and

$$\sum_{i=1}^{M} x(i) = mD$$

give checks on the data.

We thus have mD 'complete' values of W_j. However, the lack of independence makes exact calculation of the 'expected' number of gaps of size k in D draws a complex matter, so to test the UK data we simulated 160000 sets of 96 random draws to obtain the estimates of these values shown in Table 1 (the standard error of each is less than 0.02).

Tests C and D give some information on possible associations between the numbers chosen in a given draw.

2.3. *Test C: Sum of the Numbers*

Let

$$S(d) = \sum_{j=1}^{m} B_j(d).$$

Then, under hypothesis H_0, these sums have theoretical mean and variance $\mu = m(M+1)/2$ and $\sigma^2 = m(M+1)(M-m)/12$ respectively, against which we can match the sample mean and sample variance of $\{S(d){:}\ 1 \leqslant d \leqslant D\}$. Denoting the latter by U and V respectively, then the test statistics are $Z \equiv (U-\mu)/\sigma$ as standard normal, and $(D-1)V/\sigma^2$ as χ^2_{D-1}.

2.4. *Test D: Odd–Even Combinations*

Let $T(d)$ denote the number of even numbers among the m chosen in draw d. The null distribution of T is hypergeometric, and its empirical distribution can be compared with that expected. In the UK National Lottery, it will take some time (about 300 draws in total) to make the expected values in all seven cells sufficiently large to avoid combining the extreme categories {0, 1} and {5, 6} for the purposes of a conventional goodness-of-fit test.

The successive values in Fig. 1 are very highly correlated and are only independent when separated by 41 or more draws, but it is interesting that values below the theoretical mean predominate, hinting at too even a spread of numbers. The results for test D in Table 1 also have a mild suggestion that the odd–even split is insufficiently extreme, but we would expect all six numbers to have the same parity only once in 45 draws. Overall, the tests show no significant evidence of non-random behaviour.

3. DATA ON GAMBLER CHOICE

A successful model of the distribution of gambler choice must be consistent with what is known about these choices. However, very little information has been released about what choices UK gamblers are making. At various stages, it has

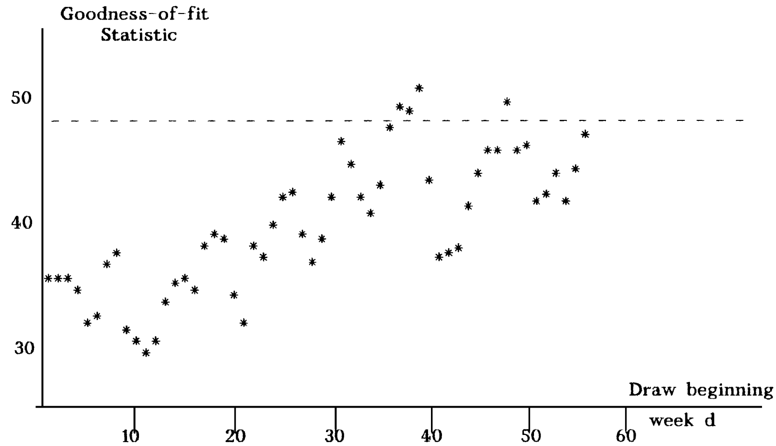

Fig. 1. Goodness-of-fit statistic *versus* draw (this follows Bellhouse (1982b) who threw doubt on the proper functioning of the machines in an early Canadian lottery; the successive χ^2-statistics, as in equation (1), are plotted for overlapping sets of 41 consecutive draws (i.e. the first point relates to draws 1–41, the next to draws 2–42, etc., the last to draws 56–96; the choice of 41 is simply to make the expected values exceed the time-honoured '5')): the long run mean value of the statistic should be 48 (- - - -)

been reported that 10000–30000 gamblers per week have selected the combination {1 2 3 4 5 6}, but, apart from the knowledge of the number of jackpot winners each draw, no other specific information seems to have been given, not even the frequencies of choice of individual numbers. In contrast, the British Columbia Lottery Corporation publishes these marginal frequencies weekly for several lotteries, including the national lotto 6/49; for example, in the two draws in the week ending October 28th, 1995, the data in *Luck* (November 6th, 1995) show the marginal frequencies of gambler choice as given in Table 2.

This ordering is very similar to that reported by Joe (1987), who gave fuller details for the draw of July 6th, 1985; we shall refer to Joe's data as draw 161. Ziemba *et al.* (1986) agreed that draw-to-draw popularity is remarkably constant, ordering the 19 least popular numbers (over 26 draws in Canada) as

40 39 20 30 41 38 42 46 29 49 48 32 10 47 1 37 28 34 45.

TABLE 2
Marginal frequencies of gambler choice

Frequency (%)	*Numbers (in order of increasing popularity)*
[1.74, 1.79]	40 39
[1.82, 1.89]	48 41 20 45 38 46 30 42
[1.90, 1.99]	37 29 36 44 49 35 47 15 34
[2.00, 2.09]	22 14 33 32 26 28 43 10 24 21 2 16 25 23
[2.10, 2.19]	18 1 19 6 13 4 12 17 8
[2.22, 2.28]	31 27 5 9 3 11
2.65	7

Space considerations inhibit the publication of the complete distribution of the 14 million or so choices, either for one draw or for a combination of several. However, some information on the distribution for lotteries is available. Stern and Cover (1989) reported on the distribution of choices among 5717817 tickets in a 6/49 Californian lottery. Zaman and Marsaglia (1990) noted that the winning combination in a Florida lottery was selected by 4676 players for the next draw, and that a further 37 previous winning combinations were each selected over 1800 times. Perhaps the most detail is given by Riedwyl (1990) who had access to the full distribution of gambler choice in week 6, 1990, of the Swiss 6/45 lottery. Here $M^* = 8\,145\,060$ and $N = 16\,862\,596$ tickets were sold, giving an average of 2.07 tickets per combination. Two combinations ({6 11 16 21 26 31} and {1 8 15 22 29 36}) were each selected more than 24000 times, and 5754 combinations were each selected more than 50 times! About 7% of sales were to the 0.07% most popular tickets.

Which combinations will prove hugely popular is plainly of great interest to gamblers. Zaman and Marsaglia (1990) and Ziemba *et al.* (1986) speculated that special dates, celebrities' birthdays, scores in sports games and current events may have transient appeal, while rows, columns or geometric patterns on the tickets may be permanent influences. The physical lay-out of some tickets is shown in Table 3. Riedwyl's data for the Swiss ticket (Table 3, part (d)) are compelling and show how important this lay-out can be. The two most popular combinations noted above are the top right–lower left and top left–lower right principal diagonals; 28 further combinations, each chosen over 2500 times, are the three other complete diagonals, the seven rows, 11 sets of six consecutive numbers in a column, {40 $\rightarrow$ 45}, {1 2 3 43 44 45}, the near diagonal {5 10 15 20 25 30}, the winning combinations from the three previous draws—and the winning combination in the last *French* lottery! Indeed, the Swiss winning combinations for the previous year

TABLE 3
Some lottery ticket lay-outs

(a) UK 6/49 ticket

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25
26	27	28	29	30
31	32	33	34	35
36	37	38	39	40
41	42	43	44	45
46	47	48	49	

(b) Lotto 6/49 (British Columbia)

	10	20	30	40
1	11	21	31	41
2	12	22	32	42
3	13	23	33	43
4	14	24	34	44
5	15	25	35	45
6	16	26	36	46
7	17	27	37	47
8	18	28	38	48
9	19	29	39	49

(c) Lotto 6/49 (Ontario)

1	8	15	22	29	36	43
2	9	16	23	30	37	44
3	10	17	24	31	38	45
4	11	18	25	32	39	46
5	12	19	26	33	40	47
6	13	20	27	34	41	48
7	14	21	28	35	42	49

(d) Swiss 6/45 ticket

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45			

were all chosen at least 295 times, and only one winning combination from the previous four years (215 draws) was chosen fewer than 59 times. The propensity of gamblers to use winning combinations is not confined simply to repeating the selection: in Riedwyl's data, the most recent winning combination was chosen 12008 times, but also the two combinations obtained by adding or subtracting 1 from each of that winning set of numbers were chosen 2342 and 1623 times respectively; similarly, those combinations differing from recent winning combinations by exactly 2, or 3 or 4 etc. (where possible) were also chosen very frequently. This transient popularity has a huge effect on a tiny proportion of the possible combinations.

The numbers of jackpot winners can be regarded as data on one randomly selected combination, on that draw. In 1211 Canadian draws, the jackpot was shared by more than 10 people only three times:

(a) draw 772 had 15 winners selecting {3 9 17 24 39 43};
(b) draw 670 had 12 winners selecting the arithmetic progression {22 27 32 37 42 47};
(c) draw 257 had 11 winners selecting {16 24 25 36 44 45}.

During the period from April 17th, 1991, to August 30th, 1995, the number of tickets per draw was reasonably stable: in 457 draws, 657 tickets won shares in jackpots. On 130 occasions there was no jackpot winner, on 141 a unique winner, 108 times two shared the jackpot and 43 times there were three sharing.

Haigh (1995) suggested that the minimum separation $m(t)$ between two successive numbers in combination t might be a useful indicator of the number of jackpot winners. Table 4 compares the actual number sharing the jackpot with that expected on a uniform distribution of gambler choice, given the numbers of tickets sold, and summed over the first 96 draws. Also included are two other criteria that may be likely to produce few jackpot winners: whether the winning combination includes several high numbers (here taken to mean at least two numbers from the final two rows of the UK ticket, Table 3, part (a)) and whether there is a large interval (here

TABLE 4

Comparison of actual and expected numbers of jackpot winners in 96 draws, according to different patterns in the winning combination

(1), criterion	*(2), frequency occurred*	*(3), actual number sharing jackpot*	*(4), expected number sharing jackpot*	*Ratio (3)/(4)*
$m(t) = 1$	47	123	230.3	0.53
$m(t) = 2$	35	117	168.2	0.70
$m(t) = 3$	5	45	23.8	1.89
$m(t) = 4$	4	150	20.8	7.23
$m(t) = 5$	3	19	14.1	1.34
$m(t) = 6$	1	7	4.65	1.50
$m(t) = 7$	1	57	4.93	11.56
Total	96	518	466.8	1.11
At least two exceed 40	30	88	151.4	0.58
Maximum separation at least 24	15	46	75.3	0.61

taken as at least 24) either below the lowest winning number, or above the highest or between two successive winning numbers. This last criterion is strongly associated with $m(t)$, both being ways of describing the evenness of the spread of numbers. Although there is a paucity of data for $m(t) \geqslant 3$, Table 4 supports the contention that gamblers have chosen their combinations too evenly spread (as Zaman and Marsaglia (1990) also noted), and with several high numbers chosen insufficiently often.

There have been several draws in Canada with no winners of the bonus prize, thereby identifying six combinations avoided by gamblers in that draw. At the other extreme, the Canadian bonus prize has been shared by more than 35 tickets only twice:

(d) draw 73, five jackpot winners and 78 bonus prizes—{1 5 7 8 11 13}, bonus 3;
(e) draw 105, one jackpot winner and 75 bonus prizes—{6 17 29 32 39 40}, bonus 47.

There is plainly some stability in choices over time. Camelot (1995) stated that 45% of players use the same numbers each week, and workplace or family lottery syndicates tend to stick with 'their' numbers. The existence of multidraw tickets, which enter given combinations for several draws, also aids this stability.

There is great draw-to-draw variation in the proportions of gamblers winning the various prizes. Since no ticket can win more than one prize, it might be expected that the correlations between the numbers of prize-winners in different categories would be near 0, or perhaps slightly negative (more match 3 winners leave fewer tickets to win match 4 prizes). Table 5 shows how wrong such speculation would be.

The positive correlations shown in Table 5 are further evidence of non-random choice by gamblers. One source of non-random choice is the use of 'wheeling' methods: a gambler confines selections to K favoured numbers ($m < K < M$) and uses a commercially produced 'wheel' to make a number of choices of m from those K numbers. Thus his selections have heavy overlap. J. A. Bather (personal communication) has calculated that so simple an association as large numbers of gamblers choosing a single number, and $m-1$ others at random, can also lead to large positive correlations in numbers of prize-winners.

TABLE 5
Correlations between the proportions of winners in the various prize categories

	Correlations for first 96 draws				*Correlations excluding weeks 9 and 70†*			
	Bonus	*Match 5*	*Match 4*	*Match 3*	*Bonus*	*Match 5*	*Match 4*	*Match 3*
Jackpot	0.77	0.86	0.62	0.43	0.38	0.66	0.47	0.33
Bonus		0.79	0.66	0.51		0.52	0.49	0.40
Match 5			0.88	0.69			0.88	0.70
Match 4				0.92				0.92

†These weeks had 133 and 57 jackpot winners, far more than in any other week, and thus heavily influencing the jackpot correlations.

Further discussion of wheeling, and the effect of the lucky dip facility, are found later, in Section 5.

4. MODELS OF GAMBLER CHOICE

The data described above come from several sources and cover many years; gambler behaviour will have varied over this period, and it seems futile to seek a single model, however complex, to account for all the features noted. But we can assess to what extent any model captures major features, and what plausible models can be discarded as grossly misleading.

4.1. *Poisson Model*

In the UK, the number of tickets sold per draw (Fig. 2) has ranged from about 48 million to 87 million (apart from the exceptional double roll-over draws 60 and 63, with over 100 million), giving a normal average μ of between 3.5 and 6.3 tickets for each of the $M^* = 13983816$ combinations. With so many gamblers and possible choices, a Poisson model has an initial attraction: independently for each draw and each combination t, the actual number choosing t follows a Poisson distribution with mean μ. However, even without the information about non-UK lotteries, the data on the numbers of UK winners, in any or all categories, reject that model emphatically. For example, there have been 17 occasions in 96 draws with no jackpot winner, as against 1–2 expected on a Poisson model; in week 9 there were 133 jackpot winners, against a mean of 5; Table 5 should show nearly zero correlations on a Poisson model; Table 4 in Moore (1997) shows that the normalized total number of prize-winners for each draw is dramatically more variable than would be consistent with a Poisson distribution.

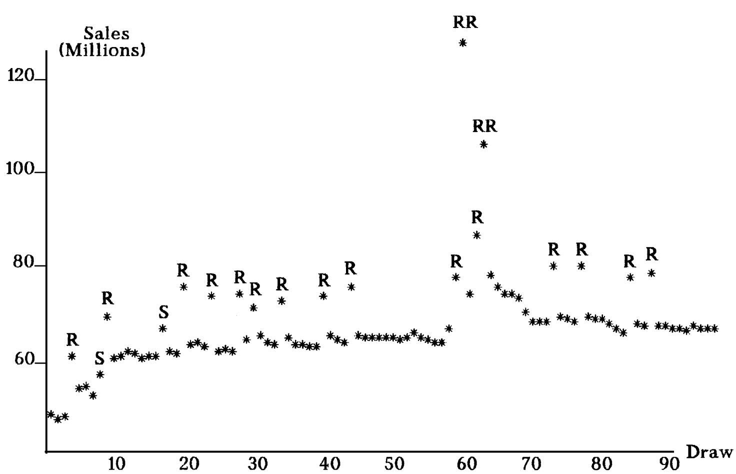

Fig. 2. Number of tickets sold each draw (R denotes a roll-over (RR is a double roll-over) and S a superdraw, i.e. an additional sum has been added to the jackpot fund)

4.2. *Models Based on Marginal Frequencies*

More sophisticated models have been offered by various researchers. In what follows, $P(t)$ denotes the probability that combination t is chosen on a single ticket, and $\{\pi_1, \pi_2, \ldots, \pi_M\}$ are the probabilities of selection of the individual numbers, with

$$\sum_i \pi_i = m.$$

Stern and Cover (1989) gave conditions under which, as total sales increase, so $P(t) \rightarrow P^*(t)$, where $P^*(t)$ maximizes the entropy, $-\Sigma\, P(t) \ln P(t)$, over all distributions $\{P(t)\}$ consistent with the marginal frequencies. This led to the multiplicative model with parameters $\theta_1, \ldots, \theta_M$, in which the probability of combination t is given by

$$P^*(t) = \prod_{i \in t} \theta_i \tag{2}$$

with the values of $\{\theta_i\}$ estimated from $\{\pi_i\}$. For draw 161, they estimated the relative frequencies of the supposed most and least popular choices, {3 7 9 11 25 27} and {20 30 39 40 41 48} respectively, finding the former four times as popular as average, and the latter a quarter as popular. However, the researchers acknowledged that the popularity of the most popular tickets is underestimated quite considerably.

Joe (1987) accepted that equation (2) might give a rough approximation to the distribution of choices but offered quite a different way of using $\{\pi_i\}$. Let X denote a distribution over the choices $\{t\}$ consistent with $\{\pi_i\}$, and let X^* denote the $\binom{M}{m}$ elements of X, arranged in non-increasing order of the frequencies; we then say that X_1 *majorizes* X_2 if the partial sums of the first j elements of X_1^* exceed those of X_2^*, for all $j \geqslant 1$. Joe suggested that the true distribution of gambler choice would be at the lower end of this majorization ordering, corresponding to near independence, given the marginal frequencies. Another criterion that he offered was to minimize $\Sigma\, P(t)^2$, subject to the marginals; for the draw 161 data, he found that this model attached zero probability to about 6900 of the possible 14 million choices. Its most popular ticket, {3 5 7 24 25 27} is about 14.5 times as popular as the average ticket.

Later, Joe (1990) suggested a class of models based on minimizing $\Sigma\, \psi\{P(t)\}$ for some convex function ψ. The choice of $\psi(u) = (u^{1+\alpha} - u)/\alpha$ for $\alpha > 0$, or its limit $\psi(u) = u \ln u$ as $\alpha \rightarrow 0$, leads to

$$P(t) = \left\{\sum_{i \in t} \theta_i\right\}_+^{1/\alpha}, \qquad \alpha > 0, \tag{3}$$

and to equation (2) as $\alpha \rightarrow 0$. Here $\{y\}_+ = \max(0, y)$, and the parameters $\{\theta_i$: $1 \leqslant i \leqslant M\}$ are to be estimated from $\{\pi_i\}$. He used the data of draw 161 to estimate $\{\theta_i\}$ for $\alpha = 0$, 0.5 and 1 in turn, and then used these estimates to predict the numbers of prize-winners in the different categories over a sample of 10 draws. These draws were selected to include some with several popular numbers, and some with several unpopular numbers. However, the predictions were usually at least two (and often more than 10) standard deviations away from the actual numbers of match 3 and match 4 prize-winners, albeit without systematic bias.

Ziemba *et al.* (1986) used the sizes of the prizes in the first 207 draws of the Canadian lottery to estimate the popularity of combinations. For each draw, the mean return for a random ticket, Y, was calculated as the weighted sum of the various prizes; the basic model is $Y = K \Pi \theta_i$ (the product being over the actual numbers drawn). The parameters $\{\theta_i\}$ were estimated via multiple regression on this model, after a logarithmic transformation. The results indicated that many combinations including the numbers 32, 29, 10, 30 and 40 would yield an average return well in excess of the stake, but that combinations including several or all of 5, 3, 13, 33, 28 and 7 would give mean returns of as little as 15–20% of the stake. These *mean* returns are associated with a large (but unquantified) standard deviation.

Zaman and Marsaglia (1990) considered several models. Their 'additive model' has

$$P(t) = M^{*-1} + K \sum_{i \in t} \left(\frac{\pi_i}{m} - \frac{1}{M} \right) \tag{4}$$

where

$$K = m \bigg/ \binom{M-2}{m-1}.$$

Unfortunately, when t consists of m numbers with small values $\{\pi_i\}$, the right-hand side of equation (4) evaluates to a negative quantity—this model demands that the marginal frequencies are more uniform than is observed. Their 'multiplicative model' repeated equation (2). Finally, they described a 'mixture model' in which particular families of combinations are assigned weights that respect $\{\pi_i\}$. For draw 161, their best fitting model gives a weight of nearly 70% to a random selection of six numbers from 49, and (for example) 3.28% to selecting $\{7\}$ together with five other numbers at random. But it is hardly better than the Poisson model at predicting numbers of prize-winners.

Finkelstein (1995) described several ways to estimate the marginal distribution of gambler choice in the Californian lotto 6/51. Suppose that $W(d)$ is the winning combination in draw d, that combination t is chosen $X(d, t)$ times and that $M(t, r)$ denotes those combinations having exactly r elements in common with t. Plainly,

$$\pi_j = \sum_{t:\, j \in t} E\{X(d, t)\} \bigg/ \sum_t E\{X(d, t)\},$$

supposed constant as d varies. The first estimate, based on the number of jackpot winners in D draws, is

$$\hat{\pi}_j = \frac{\sum_d X\{d, W(d)\}\, I\{j \in W(d)\}}{\sum_d X\{d, W(d)\}} \Bigg/ \frac{\sum_d I\{j \in W(d)\}}{mD/M}. \tag{5}$$

For any finite D, these estimates may not sum to m, but Finkelstein proved that equation (5) converges to π_j almost surely. However, with only one jackpot winner every other draw on average, convergence is very slow in practice.

His second type of estimator is based on the numbers of winners of the minor prizes. He showed that, if

$$\hat{\pi}_j(r) = \frac{\sum_d \sum_{t \in M\{W(d), r\}} X(d, t)\, I\{j \in W(d)\}}{\sum_d \sum_{t \in M\{W(d), r\}} X(d, t)} \Bigg/ \frac{\sum_d I\{j \in W(d)\}}{mD/M} \tag{6}$$

then

$$\hat{\pi}_j(r) \rightarrow \frac{Mr - m^2}{m(M-m)}\, \pi_j + \frac{m-r}{M-m} \qquad \text{almost surely} \tag{7}$$

(generalizing the previous result—put $r = m$). For his data, with $D = 176$ and $m = 6$, all the estimates (6) when $r = 3$, 4, 5 or 6 were compatible with the uniform distribution $\hat{\pi}_j = 6/51$.

His final estimator arose from assuming a multiplicative model for combination choice, based on the marginals. In draw d, any ticket would win a match 3 prize with probability $p(d)$, calculable from $\{\pi_j\}$ and $W(d)$; if $N(d)$ tickets are sold, and there are $Y(d)$ match 3 winners, let

$$S = \sum_d \{Y(d) - N(d)\, p(d)\}^2 \Big/ N(d)\, p(d)\{1 - p(d)\} \tag{8}$$

and choose estimates of $\{\pi_j\}$ to minimize S; the exact model is computationally infeasible, so the value of $p(d)$ is replaced by a good approximation; from his data, Finkelstein estimated the order of popularity as

9 3 7 8 11 6 . . . 51 43 49 48 46 50.

He noted that this suggests $\pi_i = x$ $(1 \leqslant i \leqslant 12)$, $\pi_i = y$ $(13 \leqslant i \leqslant 31)$ and $\pi_i = z$ $(32 \leqslant z \leqslant 51)$, with $12x + 19y + 20z = 6$ (the 'birthdays model' for a 6/51 lotto); this model with $x = 0.1315$, $y = 0.1196$ and $z = 0.1075$ was also consistent with the data.

Marginal frequencies are of limited help in assessing the frequency of any specific combination. Models using them tend to select the m individually most popular numbers as the most popular combination, but the researchers are well aware that there is no reason to expect this to occur. Perhaps the availability of these data from some lotteries has directed too much attention to them, at the expense of considering a combination as a whole.

4.3. *Models not using Marginal Frequencies*

Haigh (1995) suggested a possible family of models of gambler choice that we develop here. Let $\{q_j\colon j = 1, \ldots, K\}$ be any probability distribution, and let $\{C_j\colon j = 1, \ldots, K\}$ be any K non-empty subsets of the M^* combinations. Initially select C_j with probability q_j, and then choose one of the $|C_j|$ combinations, all with equal probability. Any model in this family clearly corresponds to a model in which the $\{C_j\}$ are pairwise disjoint, but perhaps at the expense of a very large value of K. However, it is sensible to insist that the union of the $\{C_j\}$ covers all possible combinations.

This family, as $\{q_j, C_j\}$ vary, is very wide. When $K = 1$ and $|C_j| = M^*$, we have the Poisson model of random choice among all combinations; with $K = M^*$ and $|C_j| = 1$

for all j, each of the M^* combinations has its own probability of selection, but there is no practical way to test this model. Joe's (1990) model (3) with $\alpha = 1$ arises by taking C_j as the combinations that include j, for $j = 1, 2, \ldots, M$. We might also base selections of $\{C_j\}$ on, for example, the following:

(a) the size of $m(t)$, the smallest difference between two consecutive values of t, so that $1 \leqslant m(t) \leqslant (M-1)/(m-1)$;
(b) for some cut-off point T $(1 < T < M)$, label numbers higher than T as high numbers; use $h(t)$, the number of high numbers in t, so that $0 \leqslant h(t) \leqslant \min(m, M-T)$ (Chernoff (1981) noted, for a different type of lottery, that numbers including 0 or 9 tended to be avoided);
(c) the maximum gap between successive members of t;
(d) whether or not 7 is included;
(e) previous winning combinations;
(f) whether all members of t are from different rows on the ticket;
(g) a birthdays model, based on the blocks $\{1 \rightarrow 12\}$, $\{13 \rightarrow 31\}$ and $\{32 \rightarrow 49\}$.

For any such model, suppose that we have the data for D draws. By analogy with equation (5), and with the same notation, one estimate of q_j is

$$\hat{q}_j = \frac{\sum_d X\{d, W(d)\}\, I\{W(d) \in C_j\}}{\sum_d X\{d, W(d)\}} \Bigg/ \frac{\sum_d I\{W(d) \in C_j\}}{|C_j| D/M^*}. \tag{9}$$

Plainly, $\hat{q}_j \rightarrow q_j$, almost surely, for $j = 1, \ldots, K$, but the estimate will only be reliable when D is sufficiently large for $|C_j| D/M^*$ to be large. For arbitrary $\{C_j\}$, the estimators (9) will not lead to such straightforward results as expression (7).

However, arguments similar to that leading to equation (8) can yield estimates of $\{q_j\}$. Suppose that $N(d)$ tickets are sold in draw d, there are $Y(d, k)$ winners of prize k, and $X(d, j)$ different combinations in C_j would win prize k. Thus the probability that a random ticket wins prize k is $p(d) = Y(d, k)/N(d)$. Our model gives this same probability as

$$q(d) = \sum q_j\, X(d, j)/|C_j|.$$

Hence, we might estimate $\{q_j\}$ for a given set $\{C_j\}$ by minimizing either

$$F(\mathbf{q}) = \sum_d \{p(d) - q(d)\}^2 \tag{10}$$

or

$$W(\mathbf{q}) = \sum_d \{Y(d, k) - N(d)\, q(d)\}^2 \Big/ N(d)\, q(d) \tag{11}$$

subject to the constraint that $\{q_j\}$ form a probability distribution.

Expression (10) gives equal weight to all draws and reduces to the linear regression of $p(d)$ on known multiples of $\{q_j\}$, subject to the equality constraint $\Sigma\, q_j = 1$, and the inequalities $q_j \geqslant 0$. Judge and Takayama (1966) described how to solve such a problem by using quadratic programming. Expression (11) is a goodness-of-fit statistic with null distribution approximately χ^2_{D-K}. Computation with equation (11) is more complicated than with equation (10) but takes account of the actual numbers of tickets and prize-winners, not just the proportions.

Broom *et al.* (1993) and Cannings *et al.* (1994) have examined the general problem of the number of local *maxima* constrained quadratic forms such as equation (10) can have. There can be at most one local minimum for equation (10) in which every $q_j > 0$, but, since our speculative model may include classes $\{C_j\}$ that should not be present, local minima with some $q_j = 0$ are relevant. Plainly, equation (11) may have several local minima in which every $q_j > 0$.

To ascertain the feasibility of estimating the parameters in such a model, given the lottery data available, we conducted a small scale experiment on a 4/12 lottery with two prize categories, in which gamblers did behave according to our model. (There were 10 classes, with considerable overlap, and probabilities ranging from 0.05 to 0.20.) Both equation (10) and equation (11) worked very well in recovering the correct model given data from 70 draws, provided that the correct 10 classes were used; the inclusion of a 'rogue' 11th class was detected, that class being given zero probability. However, if any class actually present was omitted from the model the procedures performed poorly. It seems possible that if lottery gambler behaviour does follow our model for some $\{C_j, q_j\}$, and we can identify the $\{C_j\}$ for which q_j is not too small, we can estimate their values successfully.

Encouraged by the simulation study, we sought to use equations (10) and (11) with data from the first 70 draws in the UK National Lottery, before the lucky dip option was available. To simplify matters, we combined the match 5 and bonus categories, so that 'match 5' here means simply that exactly five of the six main numbers were selected. After some experimentation, which consistently threw up certain early draws as very poor fits, we decided to omit the data from draws 1–10, when gamblers were familiarizing themselves with the game, and perhaps settling into a pattern of behaviour. We then used the data on the numbers of match 4 prize-winners in draws 11–60 to estimate $\{q_j\}$ for given $\{C_j\}$, which gives two distinct ways of judging the adequacy of the model:

(a) comparing actual with predicted match 5 and match 6 results from draws 11–60;
(b) comparing actual and predicted match 4–6 results from draws 61–70.

It has to be acknowledged that all the models tried gave poor overall fits to the lottery data. We were reasonably successful in ascertaining whether the numbers of prize-winners would be above or below average, but on several draws with very low numbers of prize-winners the estimates, although in the right direction, were far too high. For illustration, the results of fitting one model are shown in Table 6; for that model, the correlations of the actual and estimated proportions of match 4–6 winners over draws 11–60 were 0.788, 0.631 and 0.403 respectively, using equation (11). (For a Poisson model, the corresponding correlations are -0.124, 0.006 and 0.034.) Table 6 may encourage others to use these methods, with different $\{C_j\}$, in the hope of achieving a more satisfactory model. (The match 4 χ^2_{49}-value is over 75 000 from Table 6!)

TABLE 6

Best estimates of $\{q_j: 1 \leqslant j \leqslant 11\}$ obtained from equations (11) and (10), given data from draws 11–60, and the 11 classes listed

Class C	*Size* $\|C\|$	*Estimates of* q_j *using equation (11)*	*equation (10)*
All M^* combinations	13983816	0.098	0.130
$m(t) = 1$	6924764	0.064	0.029
$m(t) = 2$	3796429	0.123	0.123
$m(t) = 3$	1917719	0.060	0.050
None over 31	736281	0.128	0.129
None over 40, not all under 32	3102099	0.007	0.023
Maximum separation at most 15	5139260	0.294	0.281
All on different rows	2887500	0.099	0.101
None in outside columns	593775	0.052	0.044
Form an arithmetic progression	216	0.042	0.044
Includes 7	1712304	0.033	0.033

†$m(t)$ is the minimum separation between adjacent members of t; 'maximum separation', as in Table 4, includes the 'gap' below the lowest, or above the highest, number chosen.

5. OTHER MATTERS

On average, the National Lottery returns only 45% of money staked, but a more appropriate criterion is the comparison of the utility of the £1 stake with the utility of the added interest in the draw outcome, as well as the expected utility attached to the tiny chance of an enormous monetary prize. Note that Table 5 shows high positive correlation between the numbers of prize-winners in the various categories, and so above average jackpot or bonus prize values will be associated with above average match 5 and match 4 prizes. Thus unpopular combinations will tend to win higher prizes in all categories (except the fixed prize for match 3), and some may have a mean prize that is in excess of the stake (as Ziemba *et al.* (1986) indicated—see above). But this mean is heavily influenced by the sizes of the jackpot and bonus prizes, which will be won only once in 2 million draws; the use of the mean return on which to base decisions is questionable.

A gambler buying more than one ticket in the same draw must decide on the extent of overlap in the numbers chosen across the different tickets. Whatever the overlap, the average number of prizes depends only on the number of tickets, as does the probability of a jackpot win, provided only that all the tickets are different. Wheeling systems are sold as a means of minimizing overlap, when selections are restricted to just K of the M numbers.

Given $x < m < K \leqslant M$, a wheel identifies a collection of size R from the $\binom{K}{m}$ combinations then possible and may offer a 'guarantee' in one of the forms

(a) if the winning numbers contain x of the K selected, at least one of the R combinations contains them (sometimes weakened by the addition of 'with probability at least P%') or
(b) if the winning numbers all fall in the K selected, at least one of the R combinations contains at least x winning numbers (again, sometimes with a similarly weakening addendum).

For given K, m and x, the minimum value of R for which such guarantees can be given, and how to construct R suitable combinations, are interesting problems in combinatorial analysis: (a) is termed an (x, m) *covering* of a K-element set, and if every subset of size x belongs to exactly one of the R sets we have a *Steiner system*—see Cameron (1994), chapter 8. Of some interest is (b) when $x = 3$, $m = 6$ and $K = 49$, as this corresponds to the minimum number of tickets that must be bought to ensure at least one prize. There is a claim on the Internet lottery pages (see Section 1) that this minimum is no more than 168. (Football pools gamblers use the same ideas.)

A wheel with $K < M$ will give an increased chance of more than one prize, but a reduced chance of at least one prize, compared with the same number of tickets using all M numbers with less overlap. Thus the mean return for a given outlay from a wheeling system with K numbers chosen at random will tend to be (marginally) smaller than using all M numbers to reduce the overlap (if you win more match 4 or better prizes, the actual prize value decreases).

When the prize fund is boosted by roll-overs or superdraws, a syndicate may be tempted to seek to buy all, or nearly all, of the possible combinations. There are practical difficulties in doing so (entries of the form 'perm any 6 from 49' are not permitted), but Finkelstein (1995) reported that an Australian syndicate bought 85% of the possible tickets to win a 6/44 Virginia lottery. At the time of the first UK double roll-over, there were reports that any attempt to repeat the feat would be detected by the organizers and prevented (why?).

Fig. 2 shows how extra sales result from an enhanced jackpot. If we knew the distribution of the number of jackpot winners for a given level of sales, we could derive good estimates of the frequency of roll-overs and multiple roll-overs. But the introduction of the lucky dip facility has had a significant effect on this calculation. Data from Camelot (or the Internet) indicate that about 8% of sales are lucky dips in a normal week, but in a roll-over week about 20% of *extra* sales are lucky dips. When more lucky dip tickets are sold, it is to be expected that more of the M^* combinations are bought at least once, and so the chance of a further roll-over is reduced. Before the lucky dip was available, the attractively simple $M^*/(M^* + N)$ gave a good estimate of the roll-over probability when N tickets are sold.

The main reason why we might desire information on what choices gamblers are making is to identify unpopular combinations, in the hope of sharing the pool with fewer people. It is thus essential to note that all data relate to the past and that, as gamblers learn how others have behaved, previously popular combinations may become unpopular, and vice versa. However, the fact that 7 has been, by some distance, the most frequently selected number in the lotteries is public knowledge in Canada, and yet its popularity there remains high.

Some public excitement was generated in the UK when the Office of the National Lottery (OFLOT) instructed Camelot to change the rules, to remove the guarantee of a £10 prize for match 3, irrespective of the number of winning tickets. The purpose of the rule change was to ensure that Camelot did not have to pay out more than 45% of the proceeds, in the remote event that more than 4.5% of the tickets won a match 3 prize. Without greater knowledge of the true distribution of gambler choice, it is not possible to give a confident estimate of the chance that this will occur. But it appears far more likely that the proportion of match 3 winners will fall between 3.6% and 4.5%, which would probably lead to a match 4 prize of less than £10. Should further rule changes be made, to ensure that the values of the prizes follow their natural order?

The strategy of making a *truly* random choice—not a subjective attempt to *mimic* random selection—protects its user against falling into the same thought pattern as others, and thus considerably reduces the risk of sharing the jackpot with a large number of others. In the terminology of modern biology, ‘random choice’ is an evolutionarily stable strategy (see, for example, Maynard Smith (1982)), in the sense that, if a population as a whole uses that strategy, then no ‘invader’ using a different strategy can expect to outperform the original population. Riedwyl’s (1990) data suggest that, if there is a systematic non-random way of constructing a combination, then many gamblers are likely to use it!

Whatever methods gamblers use to choose their combinations, so long as the draw itself is random, all prize-winning combinations are chosen at random. Hence the total proportion of tickets sold that win prize C will converge, with probability 1, to N_C/M^*, where N_C is the number of winning combinations for category C.

6. CONCLUSION

The large quantities of data from the UK and similar lotteries present challenges to statisticians, but also the opportunity to help public understanding of statistics. Despite some surprise expressed in the UK when, for example, the number 39 was not drawn for over a year, and the number 44 was drawn six times running (twice as the bonus number, then four times as a main ball), statistical tests show no significant evidence that the numbers drawn are not random. But to detect a small bias, using only the data from the actual draws, may take many years. If 24 balls each had probability 10% of being selected, and the other 25 each had probability 14.4%, it would take over 600 draws for the mean value of test statistic (1) to exceed the conventional 5% significance level.

It is quite clear that gamblers’ choices of combinations have been very far from random, and Table 4 indicates certain types of combinations that have been selected much less frequently than average. As yet, no good comprehensive model of gambler choice has been found.

Commercially available ‘systems’ to play the lottery offer no advantage over the strategy of selecting numbers genuinely at random with a home-made device, but then rejecting any combination (such as 1 2 3 4 5 6, or the last winning set) that there is good reason to believe will have been chosen by many others.

ACKNOWLEDGEMENTS

Camelot and OFLOT have courteously supplied data and replied to my enquiries (I am grateful to Camelot plc for allowing me to report in Appendix A the form of the tests that are used on the numbers drawn); the Ontario Lottery Corporation sent data for 1211 draws; Harry Joe, Derek Robinson, John Bather and Robert Matthews have been very helpful. Several referees have helped me to improve the exposition, but any errors and remaining obscurities are mine.

APPENDIX A

For each of the three lottery machines used, at least 300 draws were made initially, supplemented later by data from the live draws and periodically by further test draws. Each data set is subjected to 31 statistical tests of randomness, as outlined below. The exact null distribution of any test statistic is usually calculable analytically, or by exhaustive enumeration by computer, but some null distributions were estimated by simulation. The comparison of observed values from the machines with expected values uses one or more of the Pearson goodness-of-fit statistic $\Sigma\,(O-E)^2/E$, the likelihood ratio statistic $2\,\Sigma\,O\ln(O/E)$ and the Kolmogorov–Smirnov one-sample statistic.

For consistency of notation, denote the seven numbers in the order drawn by $\{x_1, \ldots, x_7\}$, their order statistics by $y_1 < \ldots < y_7$; write $u_i = 1 +$ the integer part of $(x_i - 1)/7$, so that $1 \leqslant u_i \leqslant 7$. For each test, we describe how the test statistic is calculated. Label the tests T1–T31:

T1 (runs), $X =$ number of monotone subsequences in $\{x_i\}$, $1 \leqslant X \leqslant 6$;
T2 (up–down), each successive x-pair is either up (U) ($x_i < x_{i+1}$) or down (D), so there are 64 outcomes such as DDUDUD;
T3 (sum), $X = \Sigma\,x_i$, so $28 \leqslant X \leqslant 322$;
T4 (variance), $X = \{7\,\Sigma\,x_i^2 - (\Sigma\,x_i)^2\}/2$, so $98 \leqslant X \leqslant 12446$;
T5 (D^2), $X = (x_1 - x_5)^2 + (x_2 - x_6)^2 + (x_3 - x_7)^2$, so $3 \leqslant X \leqslant 6356$;
T6 (hypersphere), $X = \Sigma\,(x_i - 25)^2$, so $28 \leqslant X \leqslant 3619$;
T7 (evens), $X =$ number of even numbers, $0 \leqslant X \leqslant 7$;
T8–T14, X is the kth largest number ($1 \leqslant k \leqslant 7$), so $8 - k \leqslant X \leqslant 50 - k$;
T15 (poker), look on $\{u_i\}$ as a ‘poker’ hand of seven cards, and identify each hand as one of the 15 possible types, such as ‘all different’, ‘exactly two pairs’ and ‘six of a kind’;
T16 (serial correlation), X is the Pearson coefficient for the six pairs (x_i, x_{i+1});
T17 (patterns), similar to T2, except label each x_i as even (E) or odd (O); there are 128 outcomes such as EEEOEOE;
T18 (frequency), tests overall equality of frequency of all numbers drawn;
T19–T25 (position-specific frequency), as for T18, but for the numbers selected in the seven different positions, separately;
T26 (contiguity), $X =$ maximum number of contiguous values among $\{y_i\}$, so $0 \leqslant X \leqslant 7$;
T27, T28 (gaps), a draw is a ‘success’ if it has at least one number in common with the first set drawn; X is the length of the gap between successes; alternatively, the base-line draw to judge a success is the last draw that yielded a success, instead of always the first draw;
T29–T31 (coupon collecting), X is the number of draws until all possible values of $\{u_i\}$, i.e. $1 \rightarrow 7$, have been collected once (T29), twice (T30) or three times (T31).

(Tests T3, T4, T7 and T18 are, of course, the tests on all seven numbers that correspond to our tests C, D and A.) Tests T1–T26 look at aspects of randomness within one draw; the other five examine independence between draws. There is, of course, heavy dependence between the many test statistics in each of these families. How best to use all this information to decide whether the lottery machines are behaving suitably is an interesting question.

REFERENCES

Bellhouse, D. R. (1982a) Fair is fair: new rules for Canadian lotteries. *Can. Publ. Poly*, **8**, 311–320.
——— (1982b) The need for a federal lotteries review board. *Can. J. Statist.*, **10**, 213–217.
Broom, M., Cannings, C. and Vickers, G. T. (1993) On the number of local maxima of a constrained quadratic form. *Proc. R. Soc.* A, **443**, 573–584.
Camelot (1995) *Annual Report and Accounts*. Watford: Camelot.

Cameron, P. J. (1994) *Combinatorics: Topics, Techniques, Algorithms*. Cambridge: Cambridge University Press.

Cannings, C., Piff, M. J. and Vickers, G. T. (1994) Positive cones and Nash equilibria. *Proc. R. Soc.* A, **444**, 583–589.

Chernoff, H. (1981) How to beat the Massachusetts Number Game. *Math Intell.*, **3**, 166–172.

Finkelstein, M. (1995) Estimating the frequency distribution of the numbers bet on in the Californian Lottery. *Appl. Math. Comput.*, **69**, 195–207.

Haigh, J. (1995) Inferring gamblers' choice of combinations in the National Lottery. *Bull. IMA*, **31**, no. 9–10, 132–136.

Joe, H. (1987) An ordering of dependence for distribution of *k*-tuples, with applications to lotto games. *Can. J. Statist.*, **15**, 227–238.

——— (1990) A winning strategy for lotto games? *Can. J. Statist.*, **18**, 233–244.

——— (1993) Tests of uniformity for sets of lotto numbers. *Statist. Probab. Lett.*, **16**, 181–188.

Johnson, R. and Klotz, J. (1993) Estimating hot numbers and testing uniformity of the lottery. *J. Am. Statist. Ass.*, **88**, 662–668.

Judge, G. G. and Takayama, T. (1966) Inequality restrictions in regression analysis. *J. Am. Statist. Ass.*, **61**, 166–181.

Maynard Smith, J. (1982) *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Moore, P. G. (1997) The development of the UK National Lottery, 1992–96. *J. R. Statist. Soc.* A, **160**, 169–185.

Morgan, B. J. T. (1984) *The Elements of Simulation*. London: Chapman and Hall.

Riedwyl, H. (1990) *Zahlenlotto: wie Man Mehr Gewinnt*. Bern: Haupt.

Stern, H. and Cover, T. M. (1989) Maximum entropy and the lottery. *J. Am. Statist. Ass.*, **84**, 980–985.

Zaman, A. and Marsaglia, G. (1990) Random selection of subsets with specified element probabilities. *Communs Statist. Theory Meth.*, **19**, 4419–4434.

Ziemba, W. T., Brumelle, S. L., Gautier, A. and Schwartz, S. L. (1986) *Dr. Z's 6/49 Lotto Guidebook* Vancouver: Dr. Z Investments.