Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2009)
172, *Part* 4, *pp.* 813–834

Modelling the probability distribution of prize winnings in the UK National Lottery: consequences of conscious selection

R. D. Baker and I. G. McHale

University of Salford, UK

[Received September 2007. Final revision December 2008]

Summary. In the statistical and economics literature on lotteries, the problem of designing attractive games has been studied by using models in which sales are a function of the structure of prizes. Recently the prize structure has been proxied by using the moments of the prize distribution. Such modelling is a vital input into the process of designing appealing new lottery games that can generate large revenues for good causes. We show how conscious selection, the process by which lottery players choose numbers non-randomly, complicates the multivariate distribution of prize winners by introducing massive overdispersion of numbers of winners, and large correlations between the numbers of different types of prize winner. Although it is possible intuitively to reach a qualitative understanding of the data, an *a priori* model does not fit well. We therefore construct an empirical model of the joint distribution of prize winners and use it to calculate the moments of ticket value as a function of sales. The new model gives much higher estimates of ticket value moments, particularly skewness, than previously obtained. Our results will have consequences for policy decisions regarding game design. A spin-off result is that, on the basis of the results of model fitting, lottery players may increase the expected value of their ticket by strategically choosing numbers which are less popular with other lottery players.

Keywords: Conscious selection; Copula; Lotteries; Moments; Multivariate gamma; Poisson mixture

1. Introduction

Lotteries are a feature of modern life in many countries, and even those who do not play will be affected by their existence, maybe through using some sports or arts facility that has been funded through lottery monies. This makes the lottery worthy of study by statisticians and probabilists. The basic probability theory of how likely a given set of numbers is to win one of the various prizes on offer is trivial. However, the vagaries of human behaviour in playing the lottery give rise to more challenging problems in probability and statistical inference. One of these vagaries is 'conscious selection'.

Conscious selection, a term that was first coined by Cook and Clotfelter (1993), is the process by which players of a lottery do not choose their numbers randomly. Evidence of conscious selection has been found in many lotteries around the world, e.g. Cook and Clotfelter (1993) in the USA, Wang and Lin (2006) in Taiwan and various researchers in the UK (see, for example, Farrell *et al.* (2000)). No cognitive research has been undertaken to account for the existence of conscious selection and, despite no definitive answers, two plausible explanations have been

Address for correspondence: I. G. McHale, Centre for Operational Research and Applied Statistics, University of Salford, Greater Manchester, M5 4WT, UK.
E-mail: i.mchale@salford.ac.uk

0964–1998/09/172813

offered in the literature. First, players like to choose 'lucky numbers' and these lucky numbers are common across the playing population. For example, 7 has traditionally been regarded as a lucky number in the UK. Secondly, players choose numbers that are easy for them to remember, such as birthdays, and, as there is a maximum of 31 days in a month and there are only 12 months in a year, a disproportionate number of low number selections are made on entry forms. The existence of conscious selection very probably reflects a combination of these two behavioural characteristics of lottery players.

Scoggins (1995) looked for the presence of conscious selection in the Florida State Lottery by modelling the probability of the jackpot prize not being won (a 'rollover'). This probability would naively be modelled as a Poisson probability. Scoggins generalized the form of this probability, since if players do not choose numbers randomly the coverage rate, which is defined as the proportion of all possible combinations purchased, is less than would result from random selection. If the probability of winning the jackpot is p and Q tickets are sold, the rollover probability with random selection by players would be $(1-p)^Q$ (our notation). Scoggins generalized this to $(1-p)^{\alpha+\beta Q}$ where α and β are estimated from the data. From fitting to data from the Florida State Lottery, he obtained $\hat{\alpha}=8.75$ and $\hat{\beta}=0.261$, which are both statistically significantly different from the values hypothesized in the absence of conscious selection ($\alpha=0$ and $\beta=1$).

Scoggins (1995) identified the presence of conscious selection but did not measure its extent. Quantifying the extent of conscious selection is an interesting problem in itself, although the concept is somewhat arbitrary. If some or all players choose numbers and/or combinations according to some personal preferences, this does not have any effect on the distribution of winners unless their choices are correlated with each other. Thus, it could be that all buyers of lottery tickets choose their own numbers according to their particular preferences but, without correlation of preferences, the observed effects would be the same as if they chose randomly. We do not have data on the numbers that are chosen on each ticket and thus we cannot estimate the number of people with number and/or combination preferences. Nevertheless, it is desirable to obtain an index to measure the extent of conscious selection.

One possible index can be constructed by considering the observed rollover probability and comparing it with the theoretical rollover probability given no conscious selection. In the presence of conscious selection fewer number combinations are covered by the purchased tickets and thus the observed rollover probability is higher than it would otherwise be. We know the number of tickets that are sold in each draw but not the number of combinations sold. However, we can estimate, by observing the frequency of rollovers during a period, the equivalent number of tickets sold, given no conscious selection, that would produce the observed rollover probability. Effectively, this index estimates the number of identical tickets that would produce the rollover probability observed, i.e. the number of players consciously selecting with number or combination preference correlation equal to 1. Performing this calculation, for our data period, the rollover probability observed is equivalent to 24.6% of the tickets that are purchased in any one draw being identical.

Although presenting some interest as a behavioural phenomenon, the principal reason for taking the presence of conscious selection seriously is that it is relevant to the issue of designing lotteries. For instance, an important policy question is how revenue streams from lottery sales may be maximized. The issue has been addressed in various studies following Gulley and Scott (1993) in which historic sales data are employed to measure the effect on revenue when the expected value of a ticket changes. In the absence of conscious selection the resulting distribution of the numbers of winners of prizes would be Poisson, and further, given that the majority of lotteries around the world pay out several lower tier prizes in addition to the jackpot prize, the numbers of winners of each of these prizes would be independent. However, as demonstrated below, conscious selection produces an overdispersed distribution for the number of winners of any one prize, and the correlation between the numbers of winners of each of the prizes is non-zero.

Farrell *et al.* (2000) investigated the effect of conscious selection on expected value calculations. Having found strong evidence of conscious selection in the UK lotto, they presented a model for the distribution of the numbers of winners of prizes based on non-uniform number selection. Their finding is that this does not significantly affect the inferences that are drawn from models of sales as a function of the expected value of the lotto ticket. In this paper, we show that, if the model of Farrell *et al.* (2000) is used to calculate higher moments of the prize distribution, it underestimates the effects of conscious selection on the moments and we demonstrate that the variance and skewness of winnings can look very different if proper account is taken of non-random selection.

In recent developments of the sales modelling literature, researchers have increasingly used the higher moments of the distribution of winnings as explanatory variables. When the jackpot is not won, the jackpot prize money carries forward to the next draw: hence the term *rollover*. This raises the expected value for purchasers of tickets in the next draw; but, because all the extra money is typically paid into the jackpot prize, the distribution of prizes also changes. The large increases in sales that are associated with rollovers should therefore be interpreted as a response to a combination of a change in value and change in prize structure. If sales were modelled as a function of expected value and other measures representing the distribution of pay-outs for the different levels of prize, inferences could then be drawn about the effect on sales of changes in game design that produced different prize structures. Further, this would open up the possibility of making inferences from model results about the effect on revenue of any proposed changes in game design.

A natural way forward is to represent the prize structure by the higher moments of the prize distribution. Modelling sales as dependent not only on the expected value of the probability distribution of winnings of a ticket but also on the variance and skewness may reveal players' preferences for certain game design characteristics. Indeed, Garrett and Sobel (1999) found evidence from lottery games in the USA that players have a positive preference for skewness. In the UK, Walker and Young (2001) modelled sales of the UK lotto as a function of the first three moments of the prize distribution. Their results were used to compare forecast sales for two different designs of the UK lotto game, as proposed by two rival bidders for the licence to operate the lottery. It is believed that the results were influential in determining which was awarded the franchise. However, neither of these studies takes account of conscious selection, and instead both use the random selection assumption to obtain the moments of the prize distribution.

Thus there is a gap in the literature and the question remains unanswered: what is the relationship between sales and the higher moments of the prize distribution? To answer this question, we must first quantify the effect that conscious selection has on the calculation of the probability distribution of prize winnings and the moments of this distribution. It is this question that we address here.

In this paper we model the distribution of the number of winners of each prize and find the effect of this on the moments of a player's winnings from holding a lottery ticket. Using Monte Carlo simulation we then calculate the first three moments of the prize distribution and examine the effect of non-random number selection on the calculations. An intermediate result in the calculations is that we infer the popularity of the different numbers which gives a means of improving the expected value of a player's ticket by strategically choosing unpopular numbers.

The paper is structured as follows. Section 2 provides an overview of the UK lotto game, some preliminary data analysis and the equations for calculating the moments of the prize distribution. Section 3 introduces univariate modelling of the marginal distributions of the number of winners of each prize. Section 4 extends this to multivariate modelling of the joint distribution of the number of winners of each of the prizes as these joint distributions are needed to replicate properly the effect that conscious selection has on the prize distribution. Section 5 presents the results of our moments computations and Section 6 gives some closing remarks.

2. UK lotto

2.1. Game structure

The most popular of modern lottery games follow the same basic structure worldwide. Players buy tickets where they choose n numbers from a larger set of N numbers available without replacement. At a later date a draw takes place, and the winning numbers are chosen randomly from the larger set of N numbers, using a machine with numbered balls for example. The more numbers that a player matches with the balls drawn, the better the prize. If a player matches all n numbers drawn he or she wins a share of the jackpot prize. The UK lotto is a 6–49 game with five prize levels. A player chooses six numbers on his or her ticket from a possible 49. In the draw, the first six balls that are drawn are the main balls, and a seventh ball is also drawn as a bonus ball. Prizes are given for matching the first six numbers (the jackpot), five of the first six plus the bonus ball (we refer to this as 5+), five of the first six, four of the first six and three of the first six. If the jackpot is not won, it is rolled over and added to the jackpot in the next draw. This improves value for money in the 'rollover draw' as the extra prize money has been paid for by players in the earlier draw.

There are three ways in which the structure of a lottery game can vary:

(a) the choice of both n and N, where n is the number of numbers that a player must select, from a larger set of N numbers (in some recent games, players choose combinations of numbers from each of two sets of numbers rather than from just one large set);
(b) the take-out rate of the operator;
(c) the prize structure, e.g. the number of prize levels and the relative amount paid into each prize pool.

For the UK lotto the prize structure is set out in Table 1. The take-out rate is 55%, i.e. 45% of sales revenue goes into the prize fund. Players matching three balls are paid a fixed prize of £10 and funds that are required to make these pay-outs are taken from the original prize fund (45% of sales) to leave the 'prize pool'.

If lottery players chose numbers (balls) randomly, i.e. with every number being equally likely to be chosen, then the number of players matching each of three, four, five, 5+ and six balls would each be a Poisson random variable (the Poisson limit of the binomial distribution). However, as already noted, there is strong evidence that players do not choose randomly; for example low numbers are generally more popular than high numbers (Farrell *et al.*, 1999) and particular sets of numbers may be more popular because their selection forms a shape on the entry form. In the early years of the Saturday lotto, ticket sales were typically 65 million (one ticket costs £1). If players were to choose numbers randomly the expected number of jackpot winners would have been 4.65 with a standard deviation of 2.16. Yet on January 21st, 1999, there were a remarkable 133 winners of the jackpot prize. On that occasion the winning numbers formed an L-shape on the entry form and it appears that many players choose numbers on the basis of forming familiar shapes on the entry form, resulting in non-random number selection and

Table 1. Prize structure for the UK lotto

Number of matches	*Percentage of prize pool*	*Probability of winning*
6 balls	52% (plus any rollover or bonus)	7.15×10^{-8}
5+ balls	16%	4.29×10^{-7}
5 balls	10%	1.80×10^{-5}
4 balls	22%	9.69×10^{-4}
3 balls	Fixed £10 prize	0.0177

another explanation for the existence of conscious selection. The presence of conscious selection thus leads to overdispersed distributions of the numbers of winners since for some draws there are many more, or fewer, winners than would be expected under random selection.

Conscious selection, and the resultant overdispersed distribution for the number of jackpot winners, would appear to provide an explanation for the large number of winners in the South African lottery on March 3rd, 2007. On that particular occasion, the organizing body was forced to postpone the lottery by political parties calling for an investigation into the 'extremely suspect' draw in which there were nine jackpot winners. However, with conscious selection, the occurrence of nine jackpot winners would not be unduly improbable.

2.2. *Calculating the moments of the winnings from a lotto ticket*

Economists have used the moments of the winnings from a lottery ticket to aid in modelling demand for tickets for any one draw. Walker and Young (2001) presented an analytic method for calculating the first three moments of the winnings from a lotto ticket under the assumption of no conscious selection. Their method uses direct summations over the distribution of the number of winners. As a consequence of the summation, the calculation becomes computationally very expensive as sales Q increase. It is possible to use large mean approximations to inverse moments of the Poisson distribution to speed up the computations (Jones and Zhigljavsky, 2004) but, as our assumptions are more general, we use Monte Carlo simulation from the fitted distributions to calculate the moments.

Before proceeding to calculate the moments, we examine the probability distribution of the prize winnings of a lotto ticket, which is complex, and discuss a further complicating factor, the existence of lucky dip lotto tickets. First, the distribution is a mixture of discrete (£0 and £10 winners) and continuous (more than three balls matched) distributions. Second, for the continuous part, the individual prizes for matching various numbers of balls are dependent on both the number of other winners of that prize and the number of three-ball winners. In Fig. 1, the peaks resulting from the division of the jackpot between one, two, three or more winners can be seen, together with the peaks from the prizes awarded for matching 5+, five and four numbers. When the number of winners becomes very large, the resultant series of very small peaks blurs into a continuum, and it is not obvious how many modes there are. This might be an interesting question for a pure mathematician to address. Fig. 1 shows the distribution when 50 million tickets are sold and there is no rollover. It was generated by using Monte Carlo simulation. The frequencies are on a logarithmic scale (ln(frequency + 1)) and have also been scaled up by the value of the abscissa (pounds won) to enable the frequencies of large wins to be seen. In reality the frequencies of larger winnings are of course much lower than the frequencies of the smaller prizes (Table 1).

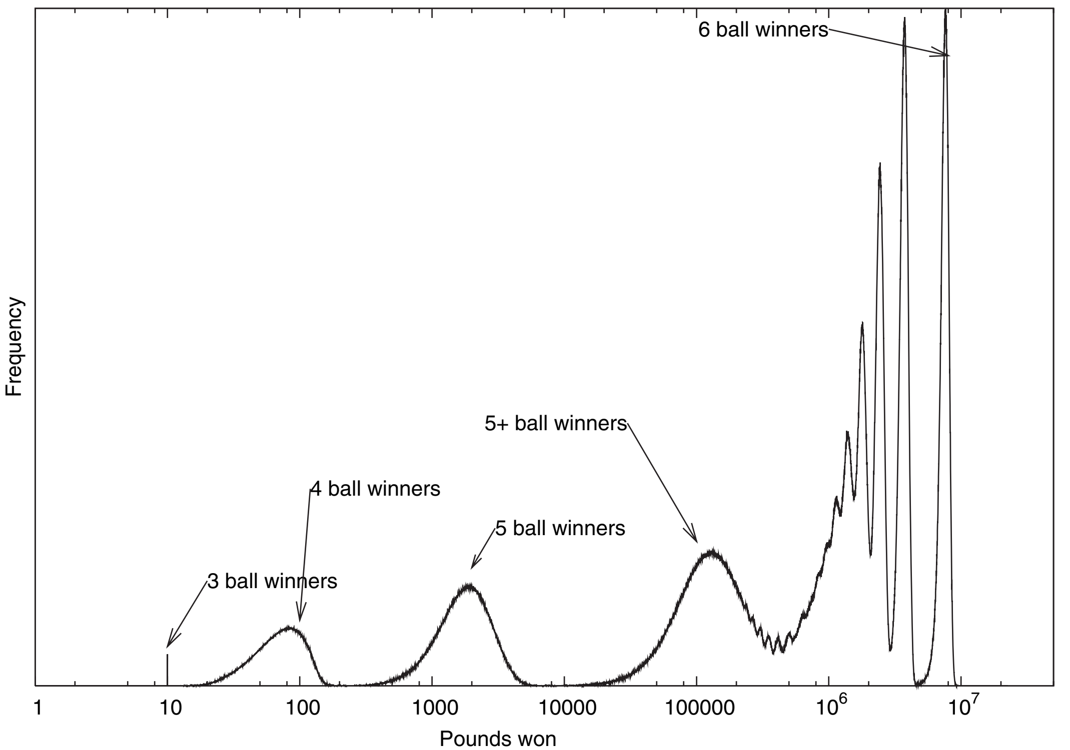


Fig. 1. Distribution of lottery winnings shown on a distorted scale: the frequencies are on a logarithmic scale (ln(frequency + 1)) and have also been scaled up by the value of the abscissa (pounds won) to enable the frequencies of large wins to be seen

A proportion of players buy a lucky dip lotto ticket. The lucky dip ticket randomly selects numbers for the player and thus eliminates any effect of conscious selection on expected ticket value. Thus, it may be that two expected values exist: one corresponding to players choosing their own numbers and exhibiting conscious selection, and a second corresponding to players whose numbers chosen are randomly selected, either by the lucky dip mechanism or by themselves. Camelot, the operator of the UK lotto, no longer publishes figures on the sales of lucky dip tickets. However, in the early days of the lucky dip, it did, and these sales were typically around 30% of total sales. Since we do not have data on the proportion of lucky dip tickets that are sold we cannot calculate the two expected values. We therefore calculate the expected value of any one ticket randomly drawn from all tickets sold. However, we suggest that a lucky dip ticket will have a higher expected value than a ticket whose numbers have been chosen with conscious selection, as the lucky dip ticket will, on average, be shared between fewer winners. The question then arises: why do players continue to choose their own numbers given that the lucky dip has a higher expected value? We can only speculate, but the answer may be twofold. First, people may believe in lucky numbers or want a set of numbers that can be remembered readily for ease of checking. Second, once a player has played with a set of numbers he or she may find it difficult to stop using them in fear that those numbers might come up in the future. This may mean that players buying extra tickets in rollovers may buy lucky dip tickets so as not to become tied in to a new set of numbers which they will feel obliged to play in subsequent weeks.

Calculation of the expected value and of other higher moments of this distribution was carried out as follows. Define the probabilities that a player correctly matches three, four, five, 5+ or six balls in a particular draw as P_3, P_4, P_5, P_{5+} and P_6 respectively. These probabilities vary as the numbers that come up are 'preferred' by players or otherwise, and are thus random variables, conditional on the winning numbers that are drawn. Their expectations are as given in Table 1. These unconditional probabilities are denoted by p_3, p_4 etc. The rth moment of the reward X is given by

$$E(X^r) = E(P_3X^r)|_3 + E(P_4X^r)|_4 + E(P_5X^r)|_5 + E(P_{5+}X^r)|_{5+} + E(P_6X^r)|_6$$

where $E(P_6X^r)|_6$ represents the expected value of the rth power of the winnings given that six balls were matched. The number of other players who achieve six matches is a random variable with mean $(Q-1)P_6$, which we take as QP_6, and from Table 1 we have that

$$E(P_6X^r)|_6 = E\left[P_6\left\{\frac{0.52(0.45Q-10n_3)+R+B}{n_6+1}\right\}^r\right]$$

where n_6 is the number of other players who matched six balls, R is the rollover and B any bonus that is added by the lottery operator. For matching five main balls and the bonus ball we obtain

$$E(P_{5+}X^r)|_{5+} = E\left\{0.16P_{5+}\left(\frac{0.45Q-10n_3}{n_{5+}+1}\right)^r\right\},$$

and so on for matching five and four balls. For matching three balls the prize is fixed and so $E(P_3X^r)|_3 = 10^r E(P_3) = 10^r p_3$.

Each expression $E(P_iX^r)|_i$ (for $i > 3$) is a function of both n_i and n_3. Thus, the marginal distributions of the n_i and their joint distributions with n_3 are needed to calculate the moments.

3. Univariate modelling

3.1. Data exploration

In this section we present the univariate modelling of the number of winners of each prize. This is needed as the distribution of the number of winners is used in the moments calculations that were described in Section 2. Four models are discussed here: the Poisson and negative binomial distributions, an *a priori* model and the Poisson–inverse gamma distribution.

We obtained publicly available data for Wednesday and Saturday lottos from Mersey World (2008) on the number of winners of each prize from inception in November 1994 to January 26th, 2008. However, the lucky dip facility was only introduced in draw 71 on March 23rd, 1996. Thus, we fit our models from then onwards so as not to overestimate the effect of conscious selection after the lucky dip was made available. We thus have a total of 1192 observations. As discussed in the previous section, the first step in the process of calculating the moments is to find the most appropriate distribution for the number of winners. Histograms for the number of winners of each prize are given in Figs 2(a) (six-ball winners) to 2(e) (three-ball winners). Also shown are the fitted negative binomial and Poisson–inverse gamma distributions (which are discussed below).

The Poisson distribution of mean μ for the distribution of the numbers of winners N for each prize would have probabilities $P_N = \mu^N \exp(-\mu)/N!$, where $\mu = aQ$, with Q sales and a the probability of winning. This distribution was fitted to the data and to maintain clarity is shown only in Figs 2(b), for 5+ winners, and 2(c), for five-ball winners, as the broken curve. The multimodal behaviour is a consequence of the fact that sales vary from week to week, and so the expected number of winners varies, whereas the Poisson distribution has a very low coefficient

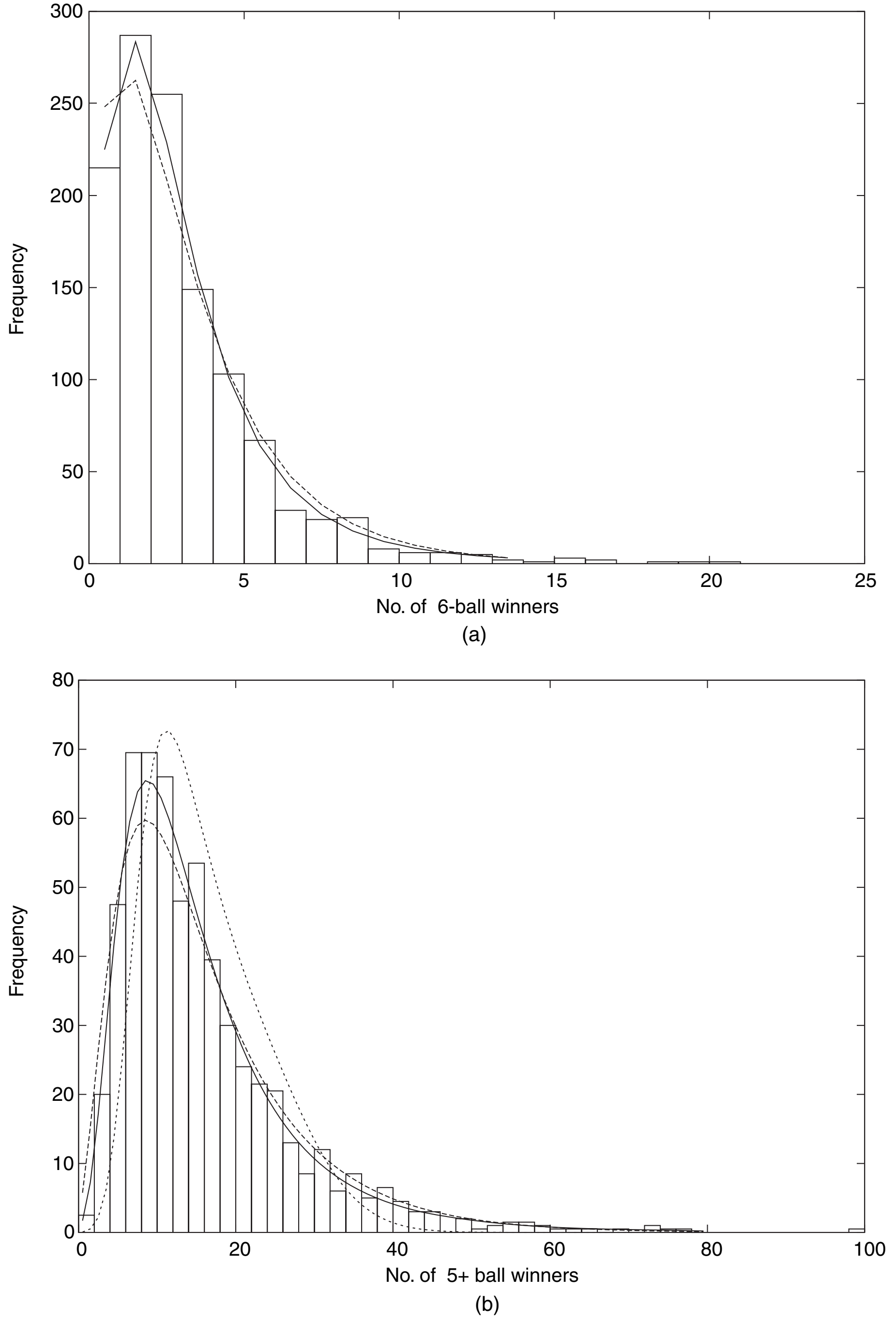
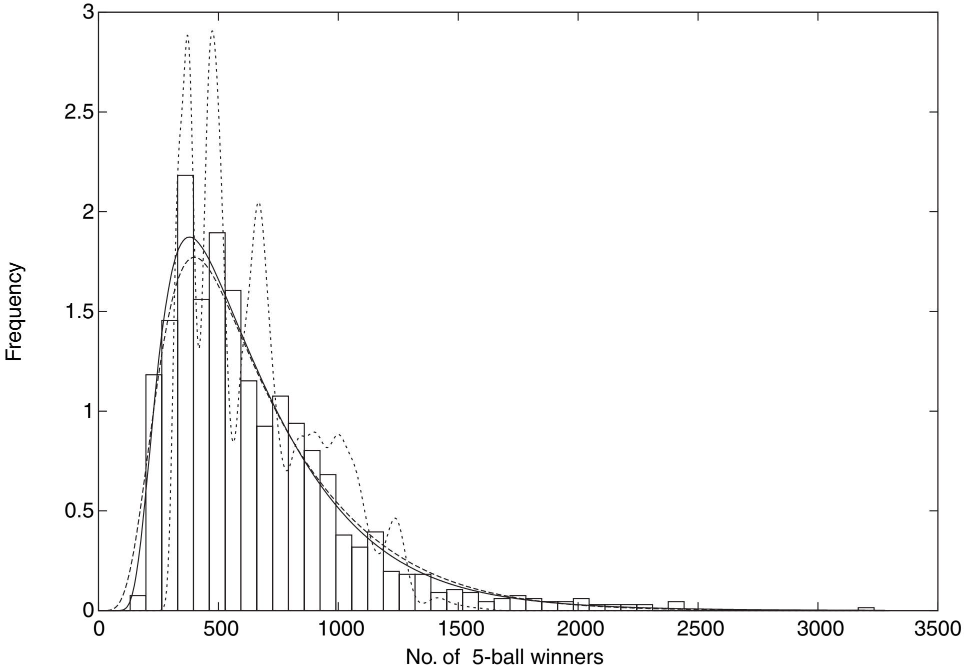
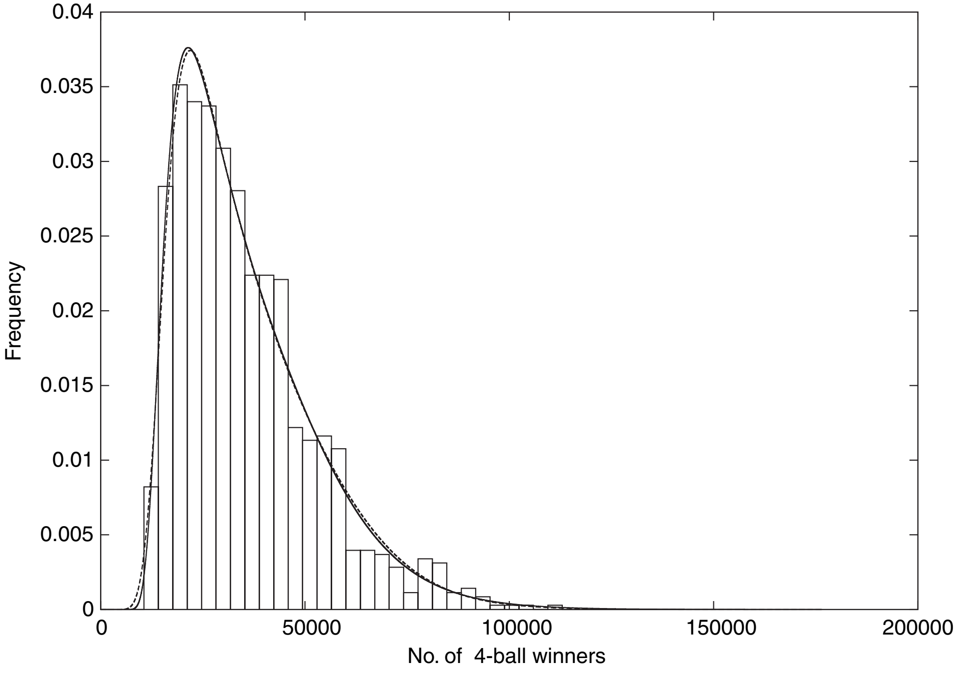


Fig. 2. (a) Histogram of six-ball winners, with fitted negative binomial (-------) and Poisson–inverse gamma (———) distributions, (b) histogram of 5+-ball winners, with fitted negative binomial (-------), Poisson (·······) and Poisson–inverse gamma (———) distributions, (c) histogram of five-ball winners, with fitted negative binomial (-------), Poisson (·······) and Poisson–inverse gamma (———) distributions, (d) histogram of four-ball winners, with fitted negative binomial (-------) and Poisson–inverse gamma (———) distributions and (e) histogram of three-ball winners, with fitted negative binomial (-------) and Poisson–inverse gamma (———) distributions



(c)



(d)

Fig. 2 (*continued*)

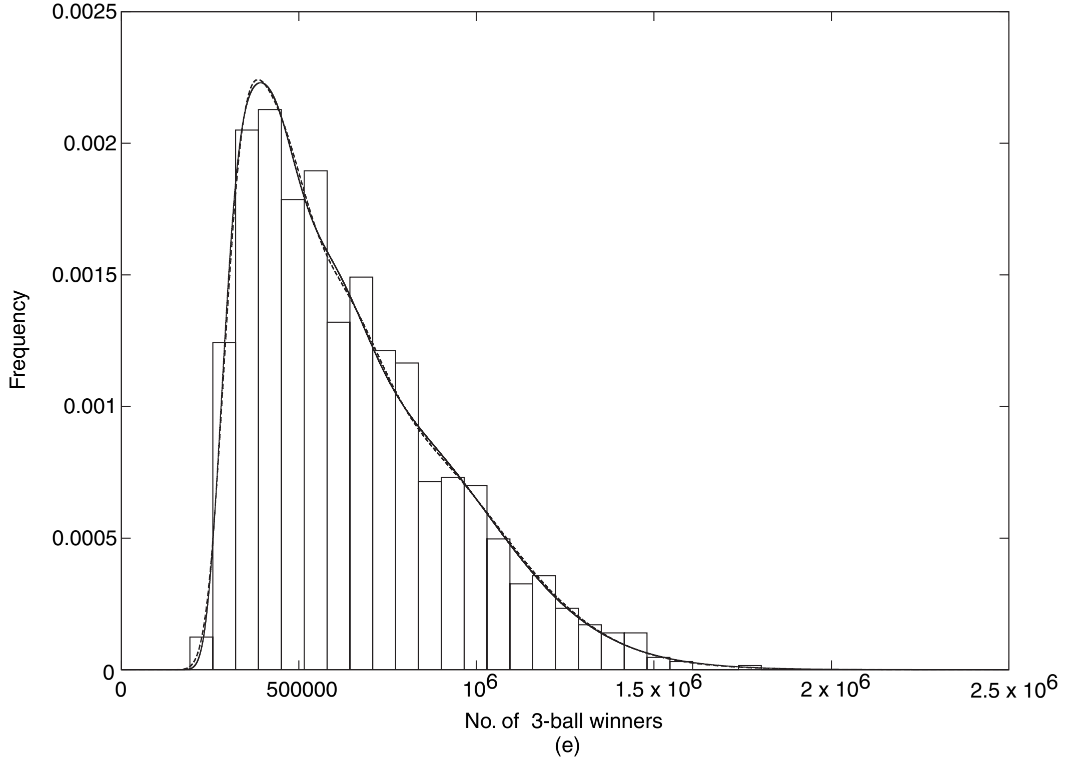


Fig. 2 (*continued*)

of variation of only $\mu^{-1/2}$. Thus, in the resulting mixture, these distributions do not overlap sufficiently to form a smooth curve. Clearly this distribution does not fit the data well. This is an important result as it means that previous studies may have been based on spurious assumptions regarding the distribution of the number of winners and this may mean that the moments that were used in these studies are not accurate. We thus seek a more appropriate fit.

An obvious modelling choice is to give a a gamma distribution, so that the probability density function (PDF) $f(a) = \alpha(\alpha a)^{\beta-1} \exp(-\alpha a)/\Gamma(\beta)$. Hence the probability P_N follows the negative binomial distribution, which is widely used for fitting overdispersed data such as numbers of industrial accidents (e.g. Johnson *et al.* (1992), chapter 5):

$$\begin{aligned} P_N &= \frac{Q^N \alpha^\beta}{N!\,\Gamma(\beta)} \int_0^\infty a^{N+\beta-1} \exp\{-(\alpha+Q)a\}\,\mathrm{d}a \\ &= \frac{(\alpha+Q)^{-(N+\beta)}\,\Gamma(N+\beta)\,Q^N \alpha^\beta}{N!\,\Gamma(\beta)} \end{aligned}$$

or

$$P_N = \frac{\Gamma(N+\beta)}{\Gamma(\beta)\,N!} \frac{(Q/\alpha)^N}{(1+Q/\alpha)^{N+\beta}}.$$

The mean is $\mu = (\beta/\alpha)Q$, which is known from the probabilities of obtaining six, five etc. balls correct, so that only one parameter need be estimated, which we take as β. As $\beta \to \infty$ the Poisson distribution is regained.

Fitted values of β are given in Table 2 with standard errors in parentheses. The increase in

Table 2. Summary of univariate model fits

Matches	χ^2 *goodness of fit for Poisson distribution*	*Results for the negative binomial distribution*		*Results for the Poisson–inverse gamma distribution*	
		$\hat{\beta}$	χ^2 *goodness of fit*	$\hat{\delta}$	χ^2 *goodness of fit*
6-ball winners	273.1 [21]	2.19 (0.17)	31.59 [13]	2.98 (0.19)	12.89 [13]
5+-ball winners	862.4 [20]	3.63 (0.18)	53.01 [28]	4.88 (0.12)	24.96 [28]
5-ball winners	1013.5 [18]	7.20 (0.29)	65.10 [28]	8.38 (0.30)	43.94 [28]
4-ball winners	756.7 [18]	17.17 (0.79)	36.02 [23]	18.03 (0.69)	33.09 [24]
3-ball winners	335.7 [18]	43.10 (1.76)	16.30 [21]	43.48 (1.73)	17.91 [22]

log-likelihood for the difference between the Poisson fit and the negative binomial is always large (well over 300). In addition, we show the χ^2 goodness-of-fit statistics. Clearly, the negative binomial distribution gives a better fit to the data in all cases, as demonstrated by both the large likelihood ratio statistic and the χ^2 goodness-of-fit statistics and corresponding degrees of freedom. Interestingly, the departure from the Poisson (random) distribution as measured by the value of β becomes smaller as the number of balls that are matched decreases. This is as might be expected from intuition and is discussed further in the next section.

The Poisson distribution is not the best fitting distribution and the data are clearly overdispersed, indicating that players do not choose numbers randomly. This is evidence of conscious selection. Indeed, if we compare the observed and predicted rollover probabilities, as Scoggins (1995) did with his model for the Florida State Lottery, the observed probability of a rollover in the lotto (no jackpot winners) is 0.180, whereas for the fitted Poisson distribution the predicted probability of a rollover is 0.0999. The difference between these probabilities is a reflection of the extent of conscious selection.

The observed and predicted rollover probabilities can be used as a goodness-of-fit measure for the three univariate models. For the negative binomial model the predicted probability is 0.2082, which is larger than the observed probability, which still suggests that a better fit might be possible. Thus we seek a model which can reproduce the observed characteristics more closely.

3.2. *An a priori model*

A simple theoretical model for conscious selection is to assume number preference, such that players choose numbers with varying probabilities $q_1 \dots q_{49}$. Since a number can be chosen only once, the q_i can be characterized as the probabilities of the first number chosen being the ith. The Poisson mean number of winners for a draw is then proportional to a random effect that is a function of the q_i. This model has 48 parameters and cannot allow for players marking out a pattern on the ticket, such as the L-shape that was described earlier, or choosing certain combinations of numbers, such as dates of birth. Nonetheless, it is worth exploring, and it was fitted to UK lottery data by Farrell *et al.* (2000). Following them, we write the probability P_6 of matching six numbers when numbers $i_1 \dots i_6$ are drawn as

$$P_6 = \sum_{\text{perms}} \frac{q_{i_1} q_{i_2} q_{i_3} q_{i_4} q_{i_5} q_{i_6}}{(1-q_{i_1})(1-q_{i_1}-q_{i_2})(1-q_{i_1}-q_{i_2}-q_{i_3})(1-q_{i_1}-q_{i_2}-q_{i_3}-q_{i_4})(1-q_{i_1}-q_{i_2}-q_{i_3}-q_{i_4}-q_{i_5})},$$

where the sum runs over the $6! = 720$ permutations of the winning numbers. This is sufficiently few for P_6 to be computed exactly.

With large sales the Poisson limit of the binomial distribution gives the log-likelihood function

$$l = \sum_{k=1}^{m} \{n_{6k} \ln(P_6 k) - Q_k P_{6k}\},$$

where, at the kth of m draws, the probability of winning is P_{6k} and sales are Q_k.

We found that the log-likelihood l could be quickly maximized by adopting a procedure that is usually rejected as inefficient: maximizing l for each of the 49 probabilities q_i in turn by using quadratic interpolation (Numerical Algorithms Group, 2008). The process converged after only a few cycles of this. The standard errors of the resulting probabilities were estimated from the profile likelihood function, and the probabilities with their error bars are plotted in Fig. 3. Some of the results that were found by Farrell *et al.* (2000) are reproduced, such as the strong preference for the number 7 (and its multiples 14 and 21), but some of their results, such as the unpopularity of the number 26, have become less pronounced. In general, higher numbers are unpopular; the Pearson correlation between the probability of choice and number was $\hat{\rho} = -0.61$.

A player wishing to choose numbers that would increase the expected value of his or her ticket would obviously choose the seven least popular numbers (35, 37, 41, 45, 46 and 49) as any prize would then be shared with fewer other winners. However, the information that is given in Fig. 3 is limited, as we have no information on how popular certain combinations of numbers are. So,

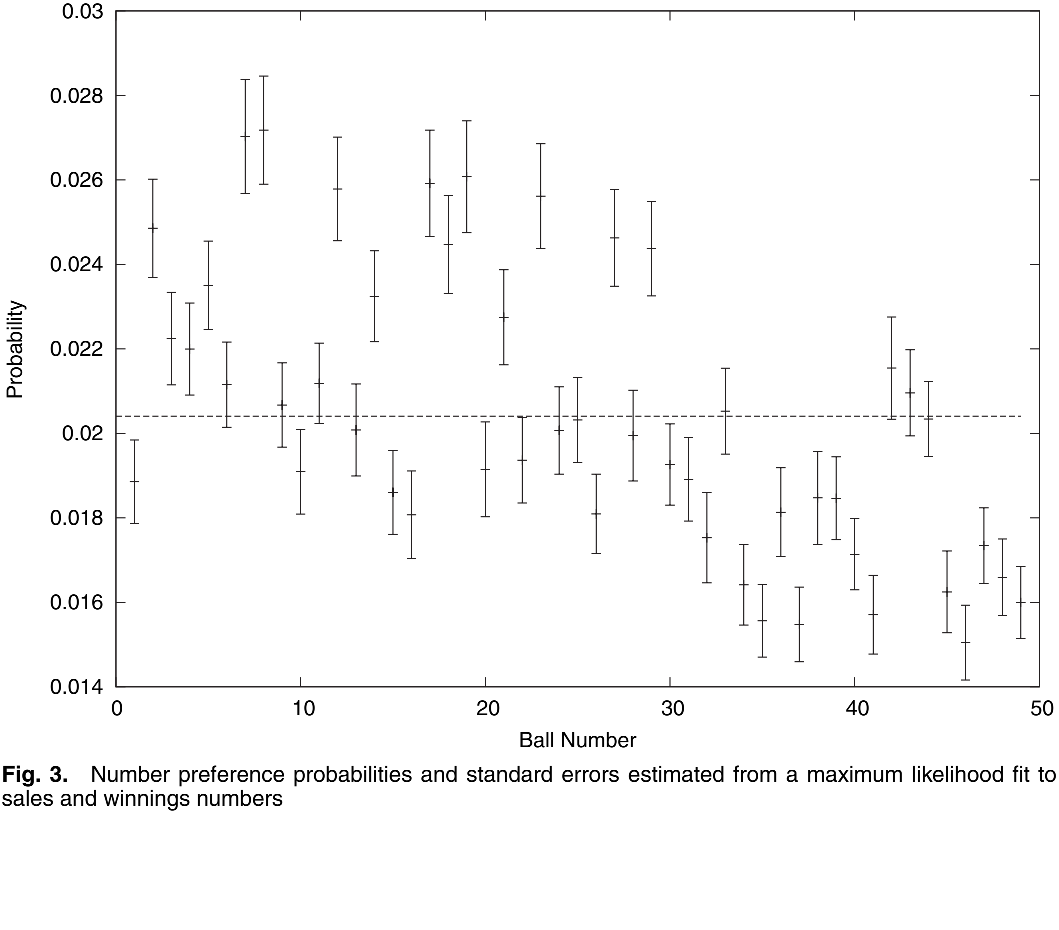


Fig. 3. Number preference probabilities and standard errors estimated from a maximum likelihood fit to sales and winnings numbers

although a player can use these results to increase the expected value of his or her ticket, they may be able to increase it further, as there may be less popular combinations than the one given. To estimate the popularity of combinations by using this method would require a model with a parameter for each of the possible combinations (nearly 14 million) and, even if a computer could handle such an optimization task, the number of observations that we have is not nearly sufficient to produce parameter estimates.

This random effect distribution gives quite a poor fit to the data. The goodness-of-fit χ^2 was 45.7 with 14 degrees of freedom, and this ignores the fact that 48 parameters were estimated from the data. The main problem is that the tail of the fitted distribution is too short. Since the dominant part of the expression for P_6 is a product of six probabilities, we would expect the central limit theorem to imply that the distribution of P_6 should be approximately log-normal. The tails of the observed distributions of the different levels of winner are, however, longer than this.

Interestingly, the correlations between the numbers of the five different levels of winner per sale are also predicted to be much too high under this model, although it is possible that this is at least partly due to the fitting of so many model parameters. Unfortunately, any attempt to reduce the noisiness of parameter estimates by shrinking the estimated probabilities towards 1/49 would shorten the tail of the predicted distribution of the number of winners still further.

We now proceed to present the results of fitting a Poisson distribution whose mean is an inverse gamma random variable. This provides an excellent fit to the data.

3.3. The Poisson–inverse gamma distribution

Karlis and Xekalaki (2005) have given a general description of Poisson mixture distributions. Several such distributions have been fitted and for brevity we present the results for the best fitting of these, namely the Poisson–inverse gamma distribution, as given in Wimmer and Altmann (1999), page 521, under the name 'Poisson reciprocal gamma' distribution. The inverse of a $\mathrm{Ga}(\gamma, \delta)$ random variable has the PDF

$$f(x) = \frac{\gamma^{\delta} x^{-1-\delta} \exp(-\gamma/x)}{\Gamma(\delta)}.$$

Making the variable X an inverse gamma random variable, where the Poisson mean number of wins is XQ, gives the Poisson–inverse gamma distribution

$$P_N = \frac{(\gamma Q)^N}{\Gamma(\delta)\, N!} \int_0^{\infty} x^{N-1-\delta} \exp\left\{-\left(\frac{1}{x} + \gamma Q x\right)\right\} \mathrm{d}x.$$

The mean is $\mu = \gamma Q/(\delta - 1)$. The constraint is that $\mu = E(P_i)Q$, with $E(P_i)$ as in Table 1. This means that γ is known in terms of δ, and δ is the only parameter to be estimated.

The probability mass function may be expressed in terms of K, the modified Bessel function of the second kind, by using an integral representation of K (Gradshteyn and Ryzhik (2007), section 8.432, number 6, page 907):

$$P_N = \frac{2(\gamma Q)^{(\delta+N)/2}}{\Gamma(\delta)\, N!} K_{N-\delta}\{2(\gamma Q)^{1/2}\}.$$

Note also that $K(z)_{-v} = K(z)_v$.

This distribution fits well: in every case better than the negative binomial distribution. See Fig. 2 and Table 2 for fitted density plots and summary model fit statistics. The predicted probability of a rollover for the Poisson–inverse gamma distribution is 0.185, which is close to the observed probability of 0.180.

4. Multivariate modelling

A further complicating factor in modelling is that, as discussed in Section 2, the marginal distributions of the number of winners of lottery prizes n_i and their joint distributions with n_3 are needed to calculate the moments. Although the univariate marginal distributions above may be adequate, a multivariate model, incorporating the dependence between the numbers of winners, is desirable. Tables 3 and 4 give the Pearson correlations between the number of winners of each prize, and the number of winners per sale for each prize. Table 4 offers a clearer insight into the relationship, because additional positive correlations between numbers of winners are induced by the variability of sales from week to week. Fig. 4 shows a scatter plot of the number of three-ball winners per million sales *versus* the number of five-ball winners per million sales.

There is clearly a strong relationship between the numbers of winners of each prize. The fact that these distributions are not independent when sales are conditioned on is a direct result of conscious selection. We first model this relationship by using arguably the best-known multivariate gamma distribution: the Mathai–Moshopoulos distribution. Next we develop our own model: an extension of the Farlie–Gumbel–Morgenstern distribution.

4.1. Mathai–Moschopoulos distribution

If possible, we would rather use a 'natural' multivariate gamma distribution than adopt the copula-based approach of gluing together marginal distributions. The distribution that was proposed by Mathai and Moschopoulos (1992) is a multivariate gamma distribution that could be used to model the Poisson random effects. It has been suggested for use in reliability, e.g. modelling the successive replacement times in a renewal process. It was chosen from the several multivariate gamma distributions that have been proposed in the literature (e.g. Kotz *et al.* (2000)) on grounds of simplicity and mathematical tractability.

Let V_i be gamma random variables with scale factor α and arbitrary shape factors. Then the model is that the inverse of the Poisson rate is a random variable $X_1 = V_1$ for six-ball winners, $X_2 = V_1 + V_2$ for 5+ winners, and so on up to $X_5 = V_1 + V_2 + V_3 + V_4 + V_5$ for three-ball winners.

Table 3. Correlations between numbers of winners

Matches	*6 balls*	*5 + bonus ball*	*5 balls*	*4 balls*
5 + bonus ball	0.493			
5 balls	0.697	0.653		
4 balls	0.529	0.614	0.932	
3 balls	0.439	0.547	0.822	0.969

Table 4. Correlations between numbers of winners per sale

Matches	*6 balls/Q*	*5 + bonus ball/Q*	*5 balls/Q*	*4 balls/Q*
5 + bonus ball/Q	0.431			
5 balls/Q	0.579	0.536		
4 balls/Q	0.486	0.460	0.904	
3 balls/Q	0.389	0.377	0.777	0.952

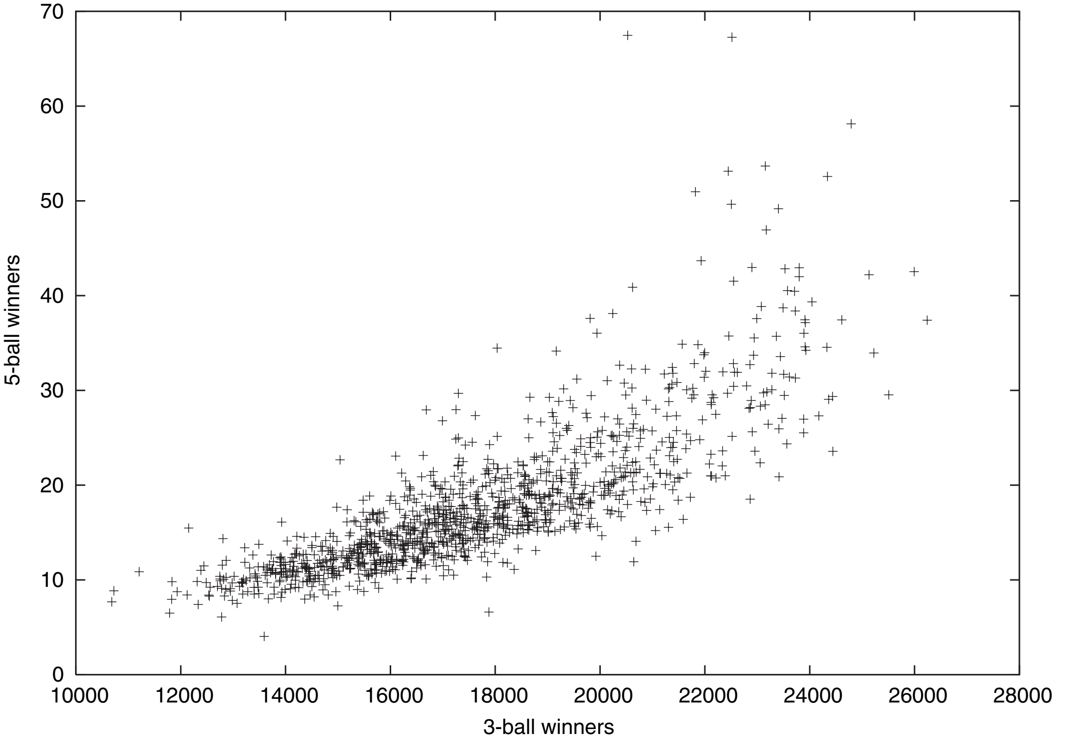


Fig. 4. Scatter plot of three-ball and five-ball winners per milllion sales

Evaluation of $P_{N_1 N_2 N_3 N_4 N_5}$ requires a five-dimensional integration. However, the means and covariance matrix can be found analytically once the marginal distributions have been fitted. Taking β_i as the ith shape parameter, we have as usual that the Poisson mean is $Q p_i$. Taking this as $c_i Q / X_i$, with $\alpha = 1$, we have that $Q p_i = c_i Q/(\beta_i - 1)$, so that $c_i = (\beta_i - 1) p_i$.

The variance of the number of winners is found by evaluating $E(N_i^2)$. This gives

$$\operatorname{var}(N_i) = \frac{c_i \bar{Q}}{\beta_i - 1} + \frac{c_i^2 \operatorname{var}(Q)}{(\beta_i - 1)(\beta_i - 2)} + \frac{(c_i \bar{Q})^2}{(\beta_i - 1)^2 (\beta_i - 2)}.$$

The variance of the number of winners per sale is

$$\operatorname{var}\left(\frac{N_i}{Q}\right) = \frac{c_i (\overline{1/Q})}{\beta_i - 1} + \frac{c_i^2}{(\beta_i - 1)^2 (\beta_i - 2)}$$

where $\overline{1/Q}$ is the mean reciprocal sales.

For the calculation of $\operatorname{cov}(N_i, N_j)$, assume that $j > i$, and let the gamma-distributed random variable that is added to X_i have shape parameter γ_{ij}, so that $\beta_j = \beta_i + \gamma_{ij}$. Then

$$E(N_i N_j) = \frac{I}{\Gamma(\beta_i)\,\Gamma(\gamma_{ij})},$$

where

$$I = \int_0^\infty \int_0^\infty \frac{x^{\beta_i - 1} y^{\gamma_{ij} - 1} \exp(-x - y)\,\mathrm{d}x\,\mathrm{d}y}{x(x+y)}.$$

To evaluate I, define $J(\alpha)$ identical to I except that the exponent is now $-\alpha(x+y)$.

Differentiating J with respect to α,

$$\frac{\mathrm{d}J}{\mathrm{d}\alpha} = -\alpha^{-\beta_i+\gamma_{ij}-1}\,\Gamma(\beta_i-1)\,\Gamma(\gamma_{ij}).$$

Integrating with respect to α, we obtain

$$I = \frac{\Gamma(\beta_i-1)\,\Gamma(\gamma_{ij})}{\beta_i+\gamma_{ij}-2}$$

or

$$E(N_i N_j) = \frac{c_i c_j \{\mathrm{var}(Q)+\bar{Q}^2\}}{(\beta_i-1)(\beta_j-2)}.$$

Finally,

$$\mathrm{cov}(N_i, N_j) = \frac{c_i c_j\,\mathrm{var}(Q)}{(\beta_i-1)(\beta_j-2)} + \frac{c_i c_j \bar{Q}^2}{(\beta_i-1)(\beta_j-1)(\beta_j-2)} \tag{1}$$

for $j>i$. The covariances for winnings per sale are derived similarly. The results for winnings and winnings per sale by using the fitted shape parameters are given in Tables 5 and 6 respectively.

The marginal fits for the Mathai–Moschopoulos distribution are good, because we preserve the Poisson–inverse gamma marginals. Once the marginal distributions have been specified, the correlations are fixed. The correlations that are thus predicted are somewhat too low, particularly those with three-ball winner numbers. This happens because this distribution is not very long tailed and so has a high β. From equation (1) this must necessarily give low covariances.

We now present an alternative model that is sufficiently flexible to generate the correlations that are required.

4.2. *A copula-based approach*

To calculate the moments of the prize winnings distribution we need a more appropriate multivariate model of the number of winners of each prize. It would have been more satisfying to

Table 5. Predicted correlations for number of winners

corr(N_i, N_j)	*6 balls*	*5 + bonus ball*	*5 balls*	*4 balls*
5 + bonus ball	0.552			
5 balls	0.445	0.731		
4 balls	0.377	0.620	0.844	
3 balls	0.341	0.560	0.763	0.904

Table 6. Predicted correlations for number of winners per sale

corr$(N_i/Q, N_j/Q)$	*6 balls*	*5 + bonus ball*	*5 balls*	*4 balls*
5 + bonus ball	0.448			
5 balls	0.316	0.587		
4 balls	0.208	0.386	0.651	
3 balls	0.129	0.240	0.405	0.621

have constructed a theoretical model of conscious selection, or, failing that, to have found a 'natural' multivariate gamma distribution for the random effect that would have fitted the data well. Since neither of these approaches succeeded, it was necessary to fall back on the use of a copula to join the well fitting Poisson–inverse gamma marginal distributions of n_i $(i>3)$ and n_3 together. We chose to model the gamma random effects for the Poisson mean numbers of winners by using a copula, rather than simply joining the Poisson–inverse gamma marginal distributions for the numbers of winners together directly. We thus attempted to construct a model that was as theoretically based as possible and was not purely descriptive.

Joe (1997) gave four criteria for a desirable multivariate model: interpretability, closure under taking margins (the lower order marginal distributions should belong to the same parametric family), a wide range of dependence and a closed form expression for the distribution function and density. The model that was introduced by Baker (2008) has these four properties and was used here. As this class generalizes the Farlie–Gumbel–Morgenstern copula (see, for example, Kotz *et al.* (2000), Drouet-Mari and Kotz (2001) and, most recently, Fischer and Klein (2007)), we call it the extended Farlie–Gumbel–Morgenstern copula.

Briefly, this class of copula has a simple interpretation in terms of order statistics. A random variable from a distribution can be generated by forming m order statistics, and randomly choosing one of them. This can be done for two or more variables, say X and Y, so that we randomly choose one of $X_{(1)}\ldots X_{(m)}$ and one of $Y_{(1)}\ldots Y_{(m)}$. If the order statistics are paired, so that we randomly choose one of the pairs $(X_{(1)},Y_{(1)})\ldots(X_{(m)},Y_{(m)})$, the marginal distributions are unchanged, but a positive correlation between the random variables is induced. The Spearman correlation is in fact $(m-1)/(m+1)$. The resulting distribution is mixed with the distribution of independent random variables to allow the correlation to be variable.

Writing the PDF and distribution function of X as $f(x)$ and $F(x)$ respectively, and correspondingly $g(y)$ and $G(y)$ for Y, the bivariate PDF $h(x,y)$ is

$$h(x,y)=(1-q)\,f(x)\,g(y)+(q/m)\sum_{k=1}^{m} f_{k,m}(x)\,g_{k,m}(y),$$

where $0\leqslant q\leqslant 1$ and

$$f_{k,m}(x)=m\binom{m-1}{k-1}F(x)^{k-1}\{1-F(x)\}^{m-k}\,f(x).$$

This has Spearman correlation $\rho_s=q(m-1)/(m+1)$.

Using this bivariate PDF for the random effects X and Y of the Poisson means QX and QY of the distributions of winners, we obtain the discrete probability $P_{N,M}\equiv\Pr(X=N,Y=M)$ as

$$P_{N,M}=(1-q)S^x_{N,1,1}S^y_{M,1,1}+(q/m)\sum_{k=1}^{m}S^x_{N,k,m}S^y_{M,k,m}, \qquad (2)$$

where

$$S^x_{N,k,m}=m\binom{m-1}{k-1}\int_0^{\infty}\frac{(Qx)^N\exp(-Qx)}{N!}\,f(x)\,F(x)^{k-1}\{1-F(x)\}^{m-k}\,\mathrm{d}x,$$

where the means are constrained to their known value as in the univariate case. In the general multivariate case, univariate integrals only are needed, so fitting such models by maximum likelihood is computationally feasible.

We point out in passing that these distributions have many attractive properties. For example, negative correlations can be accommodated by pairing order statistics in opposite order, i.e. $\{X_{(1)},Y_{(m)}\}\ldots\{X_{(m)},Y_{(1)}\}$.

On fitting the probability mass function in equation (2), with $m=75$, the bivariate distributions that are given in Table 7 were obtained.

Table 7. Summary results for fitted bivariate distributions

Matches	$\hat{q}$	ρ_{obs}	ρ_{pred}
3- and 4-ball winners	0.989	0.952	0.940
3- and 5-ball winners	0.758	0.777	0.683
3- and 5+-ball winners	0.472	0.377	0.371
3- and 6-ball winners	0.604	0.389	0.364

5. Results

Having obtained a model of the joint distribution of the numbers of winners of the various levels of prize, we now proceed to the computation of the moments of the winnings from a ticket and the modelling of sales with these moments as covariates.

5.1. The moments of the prize distribution

We use Monte Carlo simulation to calculate the moments under three alternative assumptions. First, we use independent Poisson winners for each prize. Secondly we assume the independent Poisson–inverse gamma distribution as the distribution of the number of winners, and finally we assume the dependent extended Farlie–Gumbel–Morgenstern model as described in the previous section. Only the bivariate distributions with three-ball winners needed to be modelled.

The Monte Carlo procedure was simple. The random probabilities of winning the various levels of prize in a particular draw were generated from the inverse gamma or bivariate inverse gamma distributions that were discussed earlier, and the number of winners simulated from a Poisson distribution with the appropriate mean. The characterization of the bivariate inverse gamma distribution in terms of order statistics enabled random numbers from this distribution to be readily generated. The computer program was checked by allowing rollovers to be stochastic, in which case the mean value of a lottery ticket must compute to exactly 45 p.

The results for the first three moments are shown in Fig. 5, based on 1 million simulations, for sales of 0 to 80 million. The areas of interest are those of more typical sales levels, i.e. sales when there is no rollover and sales when there is a rollover. The top and bottom deciles for these two areas actually overlap; for weeks when there is no rollover the top and bottom deciles are 19.2 million and 60.3 million sales respectively and 26.3 million to 67.1 million sales when there is a rollover. The resulting area of interest is indicated on each part of Fig. 5.

The choice of assumption has a large effect on all three moments. Assuming independent Poisson–inverse gamma distributions as opposed to Poisson distributions increases the standard deviation and skewness markedly. Obtaining a higher value for the standard deviation is as we would expect since the effect of conscious selection is to generate many more, or fewer, winners than would otherwise have been expected, meaning that prize funds are shared between more or fewer people, and hence that prize monies are more variable. In particular, variability of the number of three-ball winners will greatly increase the variability of the prize money remaining to be divided between the higher winners. Very occasionally in the simulations, no money remained after paying off the three-ball winners. This has not yet happened in practice. The skewness and standard deviation increase also as we relax the independence assumption to allow dependence between the number of winners of each prize. Again this can be understood intuitively, at least with hindsight, since there will now be some draws with few three-ball winners (leaving a relatively large prize fund). Since the correlations are positive, there will also be few winners of each of the other prizes, resulting in larger prizes for the few lucky players who do win. In addition to the increasing magnitude of the standard deviation and skewness, the qualitative behaviour is also changing. The turning point in the standard deviation curve present under the Poisson model is no longer evident in the Poisson–inverse gamma or correlated Poisson–inverse gamma models. A similar story is true for the skewness.

These results are not just of statistical interest but will have more profound consequences. For instance, the problem of designing attractive lotteries has typically been tackled in the literature by looking at player preferences for the moments of the prize distribution. In such analyses, the moments are being used as a proxy for the complex prize structure of the lottery in question. In the literature on gambling, several researchers have found evidence that bettors at the race track like skewness (see, for example, Golec and Tamarkin (1998)) and, if this is true, our analysis demonstrates how the behaviour of lottery players in consciously selecting numbers can make a game more attractive by increasing skewness. Walker and Young (2001) replicated Golec and Tamarkin's result and found that UK lottery players also had a preference for skewness. However, they did not account for conscious selection in their analysis.

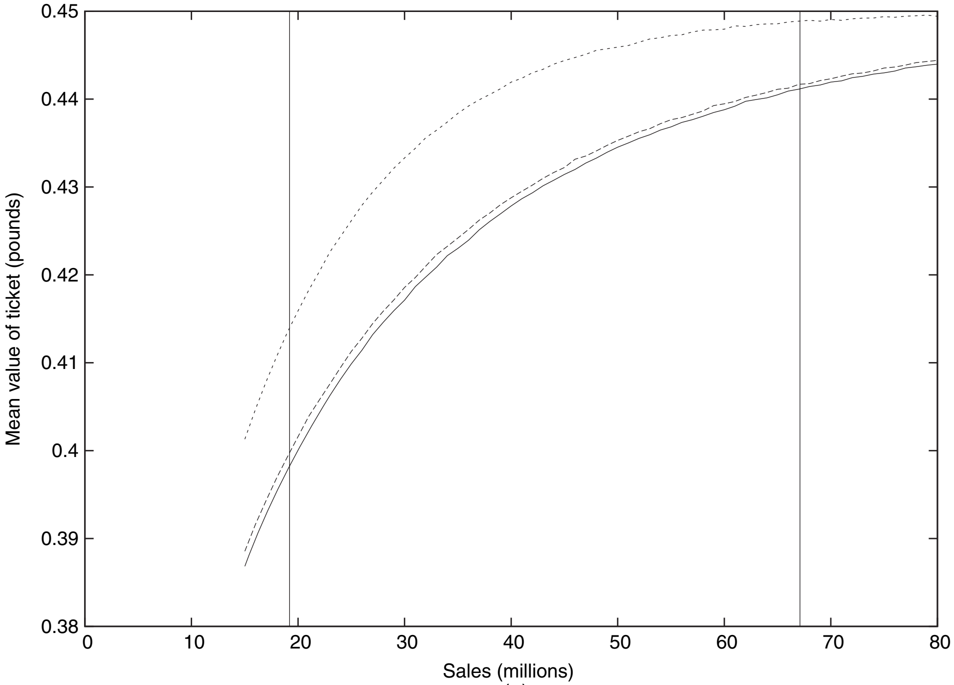
In another study of the UK lottery, Farrell *et al.* (2000) found that conscious selection has little effect on results when modelling sales. They adopted the use of an *a priori* model of conscious selection (see Section 3.2) and were concerned only with how expected value affects demand for lottery tickets. In their results, modelling sales by using the non-random-number selection distribution in comparison with the random selection distribution makes little difference to inferences made. We have seen that modelling the number of winners of prizes by using the distributions that are presented here results in considerably different values of the moments of the prize distribution, more so than when employing the *a priori* model, and such differences must be considered when designing lotteries in future.

6. Closing remarks

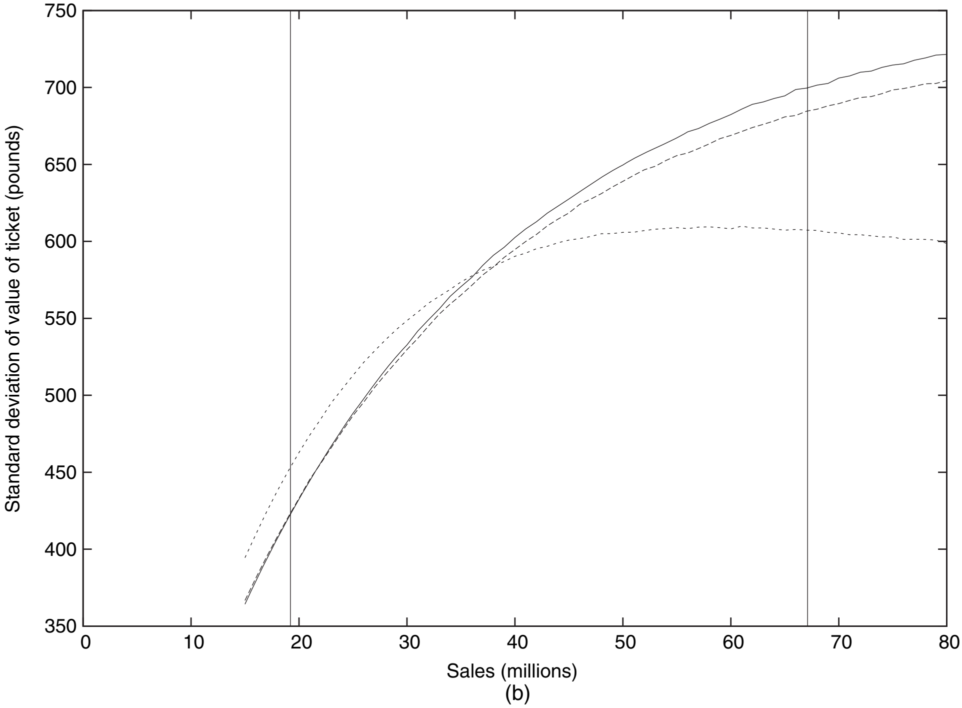
In many countries state-operated lotteries generate large revenue streams either for the state or for worthy causes such as sport, heritage or education projects. For example, since the inception of the UK National Lottery in 1994, over £22 billion has been paid into dedicated good causes funds (Department for Culture, Media and Sport, 2008). An important policy issue is how such revenue streams may be maximized.

The application of models that relate sales to mean, variance and skewness to make inferences about whether changes in game design would raise revenue is on going in the economics literature; see, for example, Forrest (2008) and Walker (2008). For the results from such work to be valid it is necessary that the moments of the probability distribution of winnings from a lottery game be calculated appropriately. In the literature to date the higher moments have been calculated on the assumption of random selection, and ignoring conscious selection of numbers.

We have seen that a realistic calculation of the moments of the prize winnings from a lottery ticket is surprisingly complicated, once conscious selection, the phenomenon of players choosing numbers non-randomly, is allowed for. The distribution that is shown in Fig. 1 must rank as one of the strangest that could be encountered in practice. The key step in calculating the moments of the prize distribution is to model the numbers of winners of each prize, which if players chose numbers randomly would simply follow the Poisson distribution. However, the tails of the observed distribution are much longer than those of the Poisson distribution and we

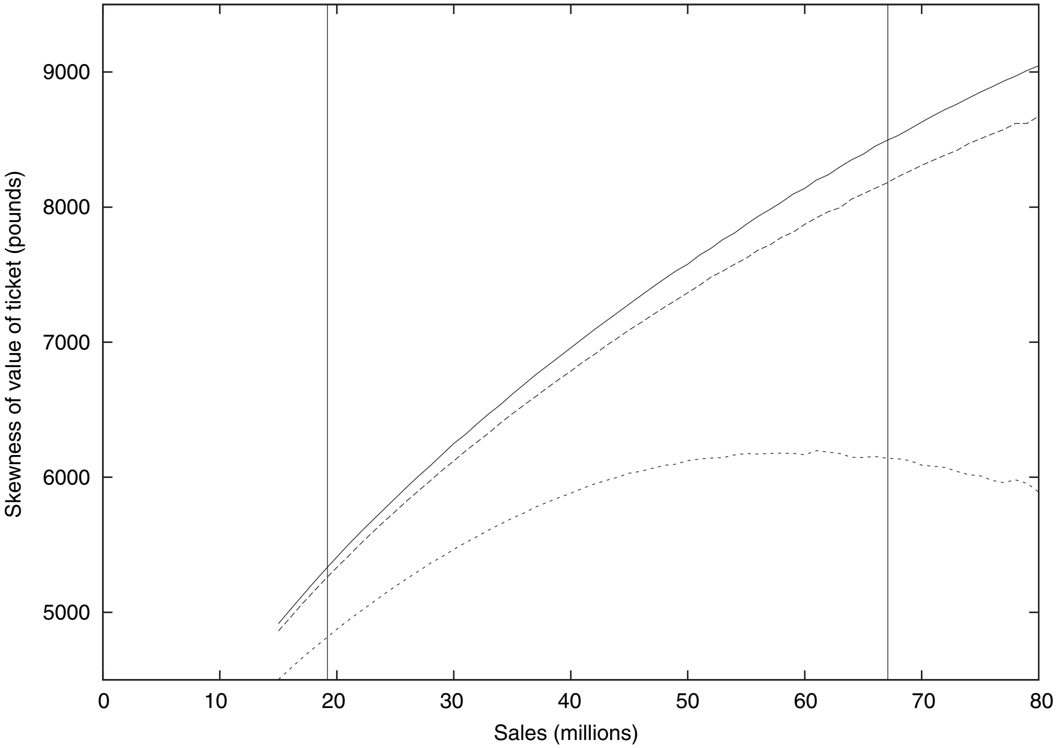


(a)



(b)

Fig. 5. (a) Mean value, (b) standard deviation and (c) skewness of a lottery ticket as a function of sales calculated under the Poisson (·······), Poisson–inverse gamma (-------) and correlated Poisson–inverse gamma (———) models: |, decile of the distribution of sales



(c)

Fig. 5 (*continued*)

found that the Poisson–inverse gamma distribution was an excellent fit to the marginal distributions of the numbers of winners of each prize. A further consequence of conscious selection is that correlations are induced between the numbers of winners of each prize, and thus we sought a multivariate model.

We generated a multivariate distribution with Poisson–inverse gamma marginals by using a copula, since there is not a multivariate generalization of the gamma distribution that allows for arbitrary correlations between the random variables. On fitting our correlated Poisson–inverse gamma distribution, we saw that the values of the moments of the prize distribution were much higher than would follow from the assumption of random-number choice. If demand models are to be used to inform the game design of lotteries then the quantitative moment calculations that were presented here will help game operators when designing new lotteries or when considering changes to existing games. Thus, when modelling the demand for lottery games one should not only use higher moments to represent prize structures but also use the appropriate probability distributions for the number of winners of each prize; otherwise inferences made may be incorrect and any conclusion that is made concerning whether or not game design changes would raise sales might be misleading.

Although we have developed a model that fits the data well, a less descriptive and more theoretical model for the multivariate distribution of the numbers of winners would be desirable. The simplified *a priori* univariate model, which was sketched out in Section 3.2, can reproduce the qualitative features of the data. However, even in the univariate case, the fit of this model proves to be unsatisfactory. Thus there is considerable scope for further work in this area.

Acknowledgements

We thank the Joint Editor, the Associate Editor, the referees and our colleague, Professor David Forrest, for their comments and suggestions on earlier versions of the paper.

References

Baker, R. D. (2008) An order-statistics-based method for constructing multivariate distributions with fixed marginals. *J. Multiv. Anal.*, **99**, 2312–2327.
Cook, P. J. and Clotfelter, C. T. (1993) The peculiar scale economies of Lotto. *Am. Econ. Rev.*, **83**, 634–643.
Department for Culture, Media and Sport (2008) General UK lottery information. Department for Culture, Media and Sport, London. (Available from `http://www.culture.gov.uk`.)
Drouet-Mari, D. and Kotz, S. (2001) *Correlation and Dependence*. London: Imperial College Press.
Farrell, L., Hartley, R., Lanot, G. and Walker, I. (2000) The demand for Lotto: the role of conscious selection. *J. Bus. Econ. Statist.*, **18**, 228–241.
Farrell, L., Morgenroth, E. and Walker, I. (1999) A time series analysis of U.K. Lottery sales: long and short run price elasticities. *Oxf. Bull. Econ. Statist.*, **61**, 513–526.
Fischer, M. and Klein, I. (2007) Constructing generalized FGM copulas by means of certain univariate distributions. *Metrika*, **65**, 243–260.
Forrest, D. (2008) Demand issues in the market for Lotto and similar games. In *Gaming in the New Market Environment* (ed. M. Viren). Basingstoke: Palgrave Macmillan.
Garrett, T. A. and Sobel, R. S. (1999) Gamblers favor skewness, not risk: further evidence from United States' lottery games. *Econ. Lett.*, **63**, 85–90.
Golec, J. and Tamarkin, M. (1998) Bettors love skewness, not risk, at the horse track. *J. Polit. Econ.*, **106**, 205–225.
Gradshteyn, I. S. and Ryzhik, I. M. (2007) *Gradshteyn and Ryzhik's Table of Integrals, Series and Products* (eds I. Jeffrey and D. Zwillinger), 7th edn. New York: Academic Press.
Gulley, O. D. and Scott, F. A. (1993) The demand for wagering on state-operated lotto games. *Natn. Tax J.*, **46**, 13–22.
Joe, H. (1997) *Multivariate Models and Dependence Concepts*. London: Chapman and Hall.
Johnson, N. L., Kotz, S. and Kemp, W. (1992) *Univariate Discrete Distributions*, 2nd edn. New York: Wiley.
Jones, C. J. and Zhigljavsky, A. A. (2004) Approximating the negative moments of the Poisson distribution. *Statist. Probab. Lett.*, **66**, 171–181.
Karlis, D. and Xekalaki, E. (2005) Mixed Poisson distributions. *Int. Statist. Rev.*, **73**, 35–58.
Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000) *Continuous Multivariate Distributions*, 2nd edn, vol. 1, *Models and Applications*. New York: Wiley.
Mathai, A. M. and Moschopoulos, P. G. (1992) A form of multivariate gamma distribution. *Ann. Inst. Statist. Math.*, **44**, 97–106.
Mersey World (2008) Lottery data archive. Mersey World. (Available from `http://lottery.merseyworld.com/`.)
Numerical Algorithms Group (2008) Routine E04ABF. In *Numerical Algorithms Group Fortran Library*. Oxford: Numerical Algorithms Group.
Scoggins, J. F. (1995) The Lotto and expected net revenue. *Natn. Tax J.*, **48**, 61–70.
Walker, I. (2008) Lottery design lessons from the Dismal Science. In *Gaming in the New Market Environment* (ed. M. Viren). Basingstoke: Palgrave Macmillan.
Walker, I. and Young, J. (2001) An economist's guide to lottery design. *Econ. J.*, **111**, F700–F722.
Wang, J.-S. and Lin, M.-Y. (2006) Number selection strategy of lottery players: an empirical study of the Taiwan Lottery. *Taiwan Econ. Forecast Poly*, **37**, 49–67.
Wimmer, G. and Altmann, G. (1999) *Thesaurus of Univariate Discrete Probability Distributions*. Essen: Stamm.