

χ^2 and the lottery

Christian Genest

Université Laval, Québec, Canada

and Richard A. Lockhart and Michael A. Stephens

Simon Fraser University, Burnaby, Canada

[Received October 2001. Revised January 2002]

Summary. The winners of many lotteries are determined by selecting at random some numbered balls from an urn. This paper discusses the use of Pearson's standard goodness-of-fit statistic to test for the equiprobability of occurrence of such lottery numbers, whether taken individually, in pairs or in larger subsets. Because the numbers are drawn without replacement, Pearson's statistic does not follow a simple χ^2 -distribution, even for large samples of draws. In fact, it can be shown that its asymptotic distribution is that of a weighted sum of χ^2 random variables. An explicit formula is given for the weights, and this result is used to check the uniformity of winning numbers in Canada's Lotto 6/49 over a period of nearly 20 years.

Keywords: χ^2 -distribution; Goodness of fit; Johnson association scheme; Lottery; Lotto 6/49; Pearson's statistic

1. Introduction

A large number of countries of the world run lotteries, and several subnational political entities (states, provinces, territories, etc.) do as well. A common way in which many of these lotteries are run is as follows. The gambler picks k distinct numbers from the set $\{1, \dots, N\}$ and then pays a fixed amount to register the selection with an authorized agent. Alternatively, the gambler may prefer to purchase a 'QuickPick' ticket whose random subset of k numbers is generated automatically by using an algorithm devised by the lottery corporation. At the close of an allotted time period, the winning combination is determined by drawing at random k balls without replacement from an urn containing identical balls labelled $1, \dots, N$. The lottery is then typically referred to as 'Lotto k/N '.

In Canada's Lotto 6/49, for example, $k = 6$ winning numbers ranging from 1 to $N = 49$ are drawn biweekly without replacement, and participants whose ticket matches three, four, five or six of the numbers drawn win prizes. Other countries such as the UK, France, Germany, Spain or the Philippines have their own weekly or biweekly draws of Lotto 6/49, and variants such as Lotto 5/26, 6/25, 6/42, 6/44, 6/45, 6/47, 6/51, 6/52, 6/53 and 6/69 have been adapted to the size of the market in Australia, Ireland and several American states; see, for example, Searle (1998).

For any lottery of the type k/N , a natural issue is whether all the numbers forming the winning combination come up with equal probability. The randomness of the selections generated by the QuickPick algorithm is of similar concern. A drawing mechanism, be it electronic or mechanical, failing this criterion would clearly induce inequity. More ambitiously, we may wish

Address for correspondence: Christian Genest, Département de Mathématiques et de Statistique, Université Laval, Québec, G1K 7P4, Canada.
E-mail: genest@mat.ulaval.ca

to test that all subsets of two numbers have the same probability of occurrence, and likewise for subsets of size 3, 4 and so on. For testing one number at a time, a natural way to proceed is to determine the frequency O_i with which the numbers $i = 1, \dots, N$ occurred in n lottery draws, and then to attempt to compare these observed counts with expected counts $E_i = nk/N$ by using the traditional Pearson statistic

$$X^2 = \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i}. \quad (1)$$

However, the asymptotic distribution of this statistic is *not* the usual χ^2 -distribution with $N - 1$ degrees of freedom, denoted by χ_{N-1}^2 , under the null hypothesis of equiprobability. This is because the observations are not drawn with replacement. Indeed, once a number has been selected among the k winning numbers drawn on a particular occasion, it cannot be chosen again in that same draw; the variability of the standard statistic (1) is thereby reduced.

In a study that was closely related to this one, Joe (1993) mentioned that it is necessary to modify X^2 to $J = (N - 1)X^2/(N - k)$ to obtain for the latter the limiting distribution χ_{N-1}^2 . The same adjustment for sampling without replacement was used by Stern and Cover (1989) in a study of the distribution of tickets purchased in Canada's Lotto 6/49. As noted by Bellhouse (1982), who had made the same point in an earlier paper on lotteries, similar problems with X^2 arise in the analysis of contingency tables based on sample survey data; see, for instance, Fellegi (1980), Holt *et al.* (1980) or Rao and Scott (1987).

It is shown below that, when testing the null hypothesis of equiprobability of subsets of size $c = 1, \dots, k$, the statistic X^2 behaves asymptotically as a sum of c independent weighted χ^2 random variables. There are then two natural courses of action: either we try to adapt Pearson's X^2 -statistic in such a way that its limiting distribution remains a simple χ^2 -distribution or we can use equation (1) and find the weights in its asymptotic distribution. Whereas Joe (1993) chose the first strategy, the second option is followed here.

The limiting distribution of X^2 in the case $c = 2$ is described in Section 2. This leads in Section 3 to a general asymptotic result including explicit formulae for the weights in the case $1 \leq c \leq k$; the details of the proof are relegated to Appendix A. Expressions for the asymptotic mean and variance of X^2 are given in Section 4, where a strategy for computing approximate limiting p -values is described. In Section 5 comparisons with Joe's statistic are made for the case $c = 2$, and in Section 6 the X^2 -statistic is used to investigate the fairness of Canada's Lotto 6/49. A short conclusion is provided in Section 7.

2. Asymptotic null distribution of X^2 in the case $c = 2$

Given a random sample of n draws of k integers among $\{1, \dots, N\}$, suppose that we wish to test that all subsets $\{i, j\}$ of numbers $1 \leq i < j \leq N$ come up with equal probability. There are $\binom{N}{2}$ such subsets, each of which corresponds to a cell in Pearson's statistic X^2 . Denoting by $O_{\{i,j\}}$ and $E_{\{i,j\}}$ the observed and expected counts for cell $\{i, j\}$ respectively, we have

$$X^2 = \sum_{i=1}^N \sum_{j=i+1}^N \frac{(O_{\{i,j\}} - E_{\{i,j\}})^2}{E_{\{i,j\}}} \quad (2)$$

with

$$E_{\{i,j\}} = e_2 \equiv n \binom{N-2}{k-2} / \binom{N}{k}$$

for all $1 \leq i < j \leq N$ under the null hypothesis of equiprobability.

Letting O and E denote the vectors of $O_{\{i,j\}}$ s and $E_{\{i,j\}}$ s listed in some order (the same for both; e.g. the lexicographic ordering $\{1, 2\}, \{1, 3\}, \dots$), we may write

$$X^2 = \left\{ \binom{N}{k} / \binom{N-2}{k-2} \right\} Y_n' Y_n$$

in terms of $Y_n = (O - E)/\sqrt{n}$, a random vector of length $\binom{N}{2}$. The null distribution of Y_n has mean 0 and covariance matrix Σ whose entries are $1/n$ times the covariances between the observed counts $O_{\{i,j\}}$ and $O_{\{i^*,j^*\}}$ for all possible choices of $1 \leq i < j \leq N$ and $1 \leq i^* < j^* \leq N$. In other words,

$$\Sigma_{\{i,j\}, \{i^*,j^*\}} = \frac{1}{n} \text{cov}(O_{\{i,j\}}, O_{\{i^*,j^*\}}).$$

Since $O - E$ is a sum of n independent and identically distributed random vectors, this covariance does not depend on n . Therefore, the asymptotic null distribution of Y_n is normal with mean 0 and covariance Σ . As the number n of draws tends to ∞ , standard results imply that X^2 converges in distribution to

$$\left\{ \binom{N}{k} / \binom{N-2}{k-2} \right\} \sum_{l=1}^{\binom{N}{2}} \lambda_l Z_l^2,$$

where the Z_l s are mutually independent standard normal variates and the λ_l s are the eigenvalues of Σ . As it happens, however, the λ_l s take only two possible non-zero values, i.e.

$$\begin{aligned} \kappa_1 &= (k-1) \binom{N-3}{k-2} / \binom{N}{k}, \\ \kappa_2 &= \binom{N-4}{k-2} / \binom{N}{k}, \end{aligned} \tag{3}$$

and these eigenvalues have multiplicity $N-1$ and $N(N-3)/2$ respectively. Consequently, the asymptotic distribution of X^2 is of the form

$$w_1 \chi_{N-1}^2 + w_2 \chi_{N(N-3)/2}^2$$

with weights $w_l = \kappa_l/e_2$, $l = 1, 2$, given explicitly by

$$\begin{aligned} w_1 &= (k-1) \binom{N-3}{k-2} / \binom{N-2}{k-2}, \\ w_2 &= \binom{N-4}{k-2} / \binom{N-2}{k-2}. \end{aligned}$$

To establish this result, first observe that, if $A_{\{i,j\}}$ denotes the event that balls i and j are among the k balls drawn, then

$$\begin{aligned} \frac{1}{n} \text{cov}(O_{\{i,j\}}, O_{\{i^*,j^*\}}) &= P(A_{\{i,j\}} \cap A_{\{i^*,j^*\}}) - P(A_{\{i,j\}}) P(A_{\{i^*,j^*\}}) \\ &= \binom{N-2-d}{k-2-d} / \binom{N}{k} - \left\{ \binom{N-2}{k-2} / \binom{N}{k} \right\}^2 \equiv r_d, \end{aligned} \tag{4}$$

where

$$d = |\{i, j\} \setminus \{i^*, j^*\}| = |\{i^*, j^*\} \setminus \{i, j\}|$$

is the cardinality of the set difference between $\{i, j\}$ and $\{i^*, j^*\}$, and hence equals 2, 1 or 0 according to whether these two sets have no, one or two element(s) in common respectively. It is thus natural to write

$$\Sigma = \lim_{n \rightarrow \infty} \{\text{var}(Y_n)\} = r_0 Q_0 + r_1 Q_1 + r_2 Q_2 \quad (5)$$

as a weighted sum of three $\{0, 1\}$ -valued matrices Q_0 , Q_1 and Q_2 of size $N(N-1)/2$ whose entries indicate which subsets $\{i, j\}$ and $\{i^*, j^*\}$ have two, one or no common element(s), respectively.

To fix ideas, consider a simpler example than the 6/49 lottery. Suppose that there are only $N = 5$ balls, $k = 3$ of which are to be drawn. Then Q_0 is the identity matrix of order $N(N-1)/2 = 10$, and $Q_0 + Q_1 + Q_2$ is a square matrix of 1s of the same size. However, the exact form of Q_1 and Q_2 depends on the selected ordering of the 10 possible subsets of size 2 of the set $\{1, \dots, 5\}$, namely $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{1, 5\}$, $\{2, 3\}$, $\{2, 4\}$, $\{2, 5\}$, $\{3, 4\}$, $\{3, 5\}$ and $\{4, 5\}$. When the subsets are ordered in precisely that fashion, we have

$$Q_1 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix},$$

and Q_2 has 1s exactly where Q_1 has 0s outside the main diagonal. Furthermore, formula (4) yields $r_0 = 21/100$, $r_1 = 1/100$ and $r_2 = -9/100$, so the matrix Σ derived from equation (5) is equal to the explicit expression (1.1) of Joe (1993).

The advantage of decomposition (5) is that the Q_d s, which are the adjacency matrices of the so-called Johnson association scheme (see, for example, Constantine (1987), page 298), may be simultaneously diagonalized and hence share the same set of eigenvectors. For instance, suppose that $\mathbf{1}$ is a vector of length $N(N-1)/2$ with all components equal to 1; this is a common eigenvector to Q_0 , Q_1 and Q_2 , since each of these three matrices has constant row sums, namely

$$Q_d \mathbf{1} = \lambda_{d,0} = \begin{cases} 1 & \text{if } d = 0, \\ 2(N-2) & \text{if } d = 1, \\ (N-2)(N-3)/2 & \text{if } d = 2. \end{cases} \quad (6)$$

Accordingly, $\mathbf{1}$ is an eigenvector of Σ whose corresponding eigenvalue equals

$$\kappa_0 = \sum_{d=0}^2 r_d \lambda_{d,0} = 0,$$

as can easily be checked by substituting expressions (4) and (6) into the above.

By use of the general results of Yamamoto *et al.* (1965) and Delsarte (1973) described in Appendix A, it may also be seen that there are $N-1$ orthogonal vectors of length $N(N-1)/2$

that are eigenvectors of Q_0 , Q_1 and Q_2 , all of which correspond to the same eigenvalue $\lambda_{d,1}$ of Q_d , namely

$$\lambda_{d,1} = \begin{cases} 1 & \text{if } d = 0, \\ N - 4 & \text{if } d = 1, \\ 3 - N & \text{if } d = 2. \end{cases}$$

Finally, the same general theory implies the existence of $N(N - 3)/2$ orthogonal eigenvectors that are common to the Q_d s, with corresponding eigenvalue

$$\lambda_{d,2} = \begin{cases} 1 & \text{if } d = 0, \\ -2 & \text{if } d = 1, \\ 1 & \text{if } d = 2. \end{cases}$$

It follows that Σ has only two strictly positive distinct eigenvalues, namely

$$\kappa_1 = \sum_{d=0}^2 r_d \lambda_{d,1}$$

and

$$\kappa_2 = \sum_{d=0}^2 r_d \lambda_{d,2},$$

which are of multiplicity $N - 1$ and $N(N - 3)/2$ respectively. The explicit expressions for κ_1 and κ_2 given in equation (3) obtain after simple calculations facilitated by the fact that $\lambda_{0,l} + \lambda_{1,l} + \lambda_{2,l} = 0$ when $l \neq 0$.

If $k = 2$, then $w_1 = w_2 = 1$ and hence X^2 is asymptotically χ^2 distributed with $\binom{N}{2} - 1$ degrees of freedom.

3. Asymptotic null distribution of X^2 in the general case

Suppose more generally that it is wished to test whether all subsets of size $c = 1, \dots, k$ are equally probable on the basis of n random lottery draws of k integers among $\{1, \dots, N\}$. If P_c denotes the collection of such subsets, the appropriate extension of statistics (1) and (2) may be written as

$$X^2 = \sum_{s \in P_c} \frac{(O_s - E_s)^2}{E_s}, \quad (7)$$

where O_s denotes the observed count for the subset $s \in P_c$ and

$$E_s = e_c \equiv n \binom{N - c}{k - c} / \binom{N}{k} \quad (8)$$

stands for the expected count for the same subset.

It is proved in Appendix A that the asymptotic distribution of X^2 as defined in equation (7) is a linear combination of c independent χ^2 random variables, i.e.

$$\sum_{l=1}^c w_l \chi_{\nu_l}^2, \quad (9)$$

where

$$w_l = \binom{k-l}{k-c} \binom{N-c-l}{k-c} / \binom{N-c}{k-c}$$

and

$$\nu_l = \binom{N}{l} - \binom{N}{l-1} = \binom{N}{l} \frac{N-2l+1}{N-l+1}. \quad (10)$$

Of course, these formulae reduce to those already presented in Section 2 when $c = 2$, and they imply that $(N-1)X^2/(N-k)$ is χ^2_{N-1} distributed when $c = 1$, as already reported by Bellhouse (1982) and Joe (1993). Also $w_1 = \dots = w_c = 1$ when $k = c$, in which case X^2 is asymptotically distributed as a χ^2 random variable with $\binom{N}{c} - 1$ degrees of freedom.

4. Asymptotic moments and computation of approximate p -values

Since the asymptotic distribution of X^2 is also that of

$$T = \left\{ \binom{N}{k} / \binom{N-c}{k-c} \right\} Y'Y$$

with Y a multivariate normal vector with mean 0 and covariance matrix Σ , a simple calculation using the fact that

$$\begin{aligned} E(Y'Y) &= \text{tr}(\Sigma) = \binom{N}{c} r_0 \\ &= \binom{N}{c} \left[\binom{N-c}{k-c} / \binom{N}{k} - \left\{ \binom{N-c}{k-c} / \binom{N}{k} \right\}^2 \right] \\ &= \left\{ \binom{N-c}{k-c} / \binom{N}{k} \right\} \left\{ \binom{N}{c} - \binom{k}{c} \right\} \end{aligned}$$

shows that the expected value of equation (7) is

$$\lim_{n \rightarrow \infty} \{E(X^2)\} = E(T) = \binom{N}{c} - \binom{k}{c}.$$

Similarly,

$$\text{var}(Y'Y) = 2 \text{tr}(\Sigma\Sigma) = 2 \binom{N}{c} \sum_{d=0}^c r_d^2 \lambda_{d,0},$$

because $\binom{N}{c} \lambda_{d,0}$ is the sum of the entries of Q_d . Using relationships (11) and (14) in Appendix A, it is then a simple matter to check that

$$\begin{aligned} \lim_{n \rightarrow \infty} \{\text{var}(X^2)\} &= \text{var}(T) = 2 \left\{ \binom{N}{c} \binom{N-c}{N-k} / \binom{N-c}{k-c}^2 \right\} \sum_{d=0}^c \binom{N-c-d}{N-k} \binom{k-c}{d} \binom{c}{d} \\ &\quad - 2 \binom{N}{c}^2 \binom{N-c}{k-c}^2 / \binom{N}{k}^2. \end{aligned}$$

Unfortunately, it does not seem possible to write the sum in a more compact form. Nevertheless, these explicit expressions for the mean and the standard deviation of X^2 can be used to compute limiting p -values from a normal approximation. However,

$$T \sim \sum_{l=1}^c w_l \chi_{\nu_l}^2$$

has an asymmetric distribution, and a better approximation for $P(T > t)$ can be found by computing $P(S > t)$ for $S = a + b\chi_{\nu}^2$, where the scalars a and b and the degrees of freedom ν are chosen so that the first three moments (or equivalently the first three cumulants) of S match those of T . In the present case, this amounts to solving

$$E(T) = \sum_{l=1}^c w_l \nu_l = a + b\nu,$$

$$\text{var}(T) = 2 \sum_{l=1}^c w_l^2 \nu_l = 2b^2 \nu$$

and

$$8 \sum_{l=1}^c w_l^3 \nu_l = 8b^3 \nu,$$

which yields

$$b = \frac{\sum_{l=1}^c w_l^3 \nu_l}{\sum_{l=1}^c w_l^2 \nu_l},$$

$$\nu = \left(\sum_{l=1}^c w_l^2 \nu_l \right)^3 / \left(\sum_{l=1}^c w_l^3 \nu_l \right)^2$$

and

$$a = \sum_{l=1}^c w_l \nu_l - b\nu.$$

In practice, the approximation

$$P(T > t) \approx P(a + b\chi_{\nu}^2 > t)$$

with the above values of a , b and ν proves very good in the upper tail, i.e. for values of t corresponding to p -values that are lower than 50%.

5. Comparison with the statistic of Joe (1993)

As mentioned in Section 1, an alternative way to test for uniformity of sets of lotto numbers was investigated by Joe (1993), who replaced the statistic $X^2 = (O - E)'(O - E)/e_c$ by another statistic J whose limiting distribution is χ^2 with $\binom{N}{c} - 1$ degrees of freedom. The general form of his statistic is

$$J = (O - E)' \Sigma^- (O - E)/n,$$

where Σ^- stands for a generalized inverse of Σ . In the case $c = 2$, we find that

$$J = \frac{b_0}{n} (O - E)'(O - E) + \frac{b_1}{n} (O - E)' Q_1 (O - E)$$

$$= \frac{b_0 e_2}{n} X^2 + \frac{b_1}{n} (O - E)' Q_1 (O - E),$$

where $b_0 = \{(k-1)N - 5k + 7\}B$ and $b_1 = -(k-2)B$ with

$$B = \frac{N(N-1)(N-2)}{k(k-1)^2(N-k)(N-k-1)}.$$

For the 6/49 lottery, for example, $b_0 = 222B$ and $b_1 = -4B$ with $B = 47 \times 28 / (75 \times 43) \approx 0.408$, and Joe's statistic can be compared, for sufficiently large n , with a χ^2_{1175} -distribution. In contrast, the simpler statistic X^2 as defined in equation (2) is asymptotically distributed as

$$\frac{215}{47} \chi^2_{48} + \frac{903}{1081} \chi^2_{1127},$$

with a mean of 1161 that is roughly the same as that of Joe's statistic (equal to 1175), but a variance of 3581.69 compared with 2350.

What really matters, however, is the relative power of X^2 and J as test statistics along credible sets of alternatives. To make such a comparison, assume that the propensity p_i that ball i is drawn depends only on its intrinsic physical properties, such as its weight. Given a set D of balls in the urn, the probability that ball i is selected would then be proportional to p_i when $i \in D$. Thus if there are N balls to start with, and if D_r denotes the set of balls remaining in the urn after $r = 0, \dots, 5$ balls have been selected, the probability of the event $B_{i,r+1}$, that the $(r+1)$ th ball selected is ball number i , would be given by

$$P(B_{i,r+1}|D_r) = \begin{cases} p_i / \sum_{j \in D_r} p_j & \text{if } i \in D_r, \\ 0 & \text{otherwise.} \end{cases}$$

Unless $p_1 = \dots = p_N$, this model provides a reasonable alternative against which to compare X^2 and J in practice. Furthermore, a smooth interpolation between this model and the null hypothesis of uniformity can be obtained by replacing each occurrence of p_i in the above formula by p_i^α with α running from 0 to 1.

As an illustration, Fig. 1 shows the power function of the statistics X^2 and J for this family of alternatives when $p_1, \dots, p_{49}$ are the observed frequencies of the balls in the first $n = 1798$ draws of Canada's Lotto 6/49. The curves are based on 10000 Monte Carlo repetitions of the same number of draws from the alternative model for various values of α . Given the number of replicates, the vertical standard error in each of the curves is at most 1/200.

For $c = 1$, the statistics X^2 and J are equivalent. As can be seen from Fig. 1, these statistics are quite powerful for this realistic set of alternatives. When $c = 2$, the power drops for both statistics but X^2 is considerably more powerful than J . This pattern may be expected to continue for larger values of c and similar types of alternative. Of course, alternatives could also be found for which Joe's statistic is more powerful than X^2 .

6. Fairness of Canada's Lotto 6/49

Lotto 6/49 was introduced in Canada in 1982 as a national lottery offering currently a weekly grand prize of at least \$2 million (currently worth about US \$1.3 million or 1.54 million euros) payable in one tax-free lump sum shared equally between ticket holders who selected the appropriate combination of six numbers from among integers 1 to 49. Depending on sales and the number of consecutive draws without a grand prize winner, the jackpot can reach up to \$20 million. Smaller prizes are available for players whose ticket matches three, four or five of the numbers drawn. A seventh 'bonus number' is drawn without replacement and it is also possible

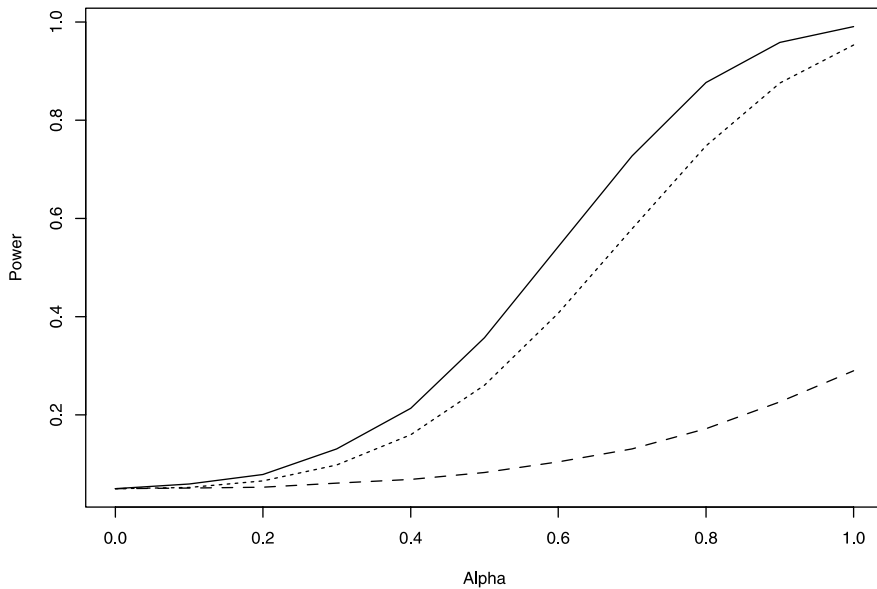


Fig. 1. Power functions for the statistics X^2 and J (—) when $c = 1$, and for the statistics X^2 (-----) and J (- - -) when $c = 2$, for the continuous set of alternatives as defined in the text

to win a prize by matching five of the first six balls and this bonus number. Very similar lotteries exist elsewhere, but Canada's has been one of the longest running; it thus provides ample data on which to apply the tests above.

The first weekly draw of Canada's Lotto 6/49 was made on June 12th, 1982, and beginning in September 1986 winning combinations have been selected twice a week, on Wednesdays and Saturdays. Results are posted as soon as they become available on provincial lottery board Web sites, e.g. British Columbia's Lottery Corporation site <http://www.bclc.com/> or Loto-Québec's French language equivalent, located at <http://www.loto-quebec.qc.ca/>. Several privately run sites also exist that provide comprehensive databases including complete historical records of the draws, up-to-date frequency charts and various statistical tools to help players to spot 'hot numbers' or 'numbers that are due up' by checking on the past occurrence of arbitrary combinations of from one to six numbers since the lottery's inception. Interested readers may refer, among others, to

<http://www.lottohome.com/>
<http://webhome.idirect.com/~forward/>
<http://www.lotterycanada.com/lottery/>

To determine whether Canada's Lotto 6/49 is fair, the results of the first $n = 1798$ draws were extracted from these sites. Fig. 2 displays the observed variability in the occurrence of the various numbers in the six-ball winning combination. The minimum and maximum observed frequencies were 193 and 269, corresponding to balls 48 and 31 respectively.

Table 1 shows the result of testing for equidistribution using statistic (7). For each possible subset size $c = 1, \dots, 6$, Table 1 gives e_c , the expected count of all possible subsets s of size c , the value of the statistic X^2 and the associated p -value based on the asymptotic distribution (9). It must be recognized that the dependence between the test statistics X^2 corresponding to

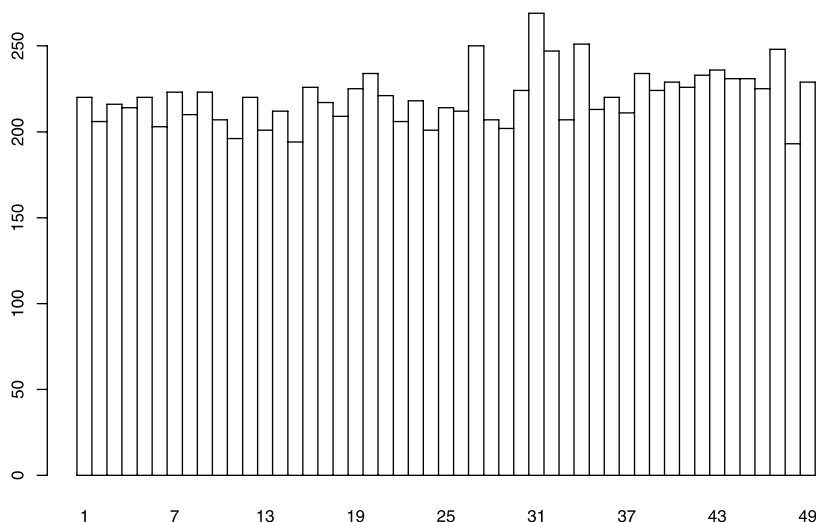


Fig. 2. Observed frequency of occurrence of balls 1–49 in the six-number winning combination of the first $n = 1798$ draws of Canada's Lotto 6/49

Table 1. Test of equidistribution for subsets of $c = 1, \dots, 6$ balls in Canada's Lotto 6/49 using the first 1798 draws spanning June 12th, 1982, and April 14th, 2001

c	e_c	X^2	p -value
1	220.1633	54.34	0.104
2	22.9337	1190.95	0.300
3	1.9518	18416.4	0.476
4	0.1273	211899.2	0.479
5	0.0056	1906878	0.534
6	0.0001	13982018	0.633

different values of c makes a multiple-comparison adjustment difficult. Nevertheless, it is worth noting that the hypothesis of equidistribution cannot be rejected at the 5% level for any value of c . In fact, even the statistic X^2 for individual numbers ($c = 1$) is not significant at the 10% level, suggesting that the drawing mechanism that is used is fair.

A slightly more nuanced conclusion emerges when all seven numbers, including the bonus number, are regarded as the basic set and subsets are tested for uniformity as above. The results are in Table 2, which gives the value of statistic (7) when $k = 7$ and $N = 49$. In the case $c = 1$, the observed value of X^2 corresponds to a p -value of 0.044 which, the multiple-comparison issue aside, would lead to the rejection of the hypothesis of equidistribution at the 5% significance level. Note, however, that this test does not take into account the order in which the balls are drawn, though this order is important in determining prize winners.

In Tables 1 and 2, the p -values given by the limiting distribution were first found by the method of Imhof (1961). The accuracy of this numerical integration technique, which can be very slow in the upper tail, was set to be within 10^{-7} . The three-moment χ^2 -approximation to the

Table 2. Test of equidistribution for subsets of $c = 1, \dots, 7$ balls in Canada's Lotto 6/49 using the first 1798 draws spanning June 12th, 1982, and April 14th, 2001

c	e_c	X^2	p -value
1	256.85714	57.64	0.044
2	32.10714	1218.06	0.164
3	3.41565	18487.51	0.357
4	0.29701	212471.8	0.238
5	0.01980	1906599	0.544
6	0.00090	13983809	0.853
7	0.00002	85898786	0.555

limiting p -values was found to be exact to the fourth decimal place and, if $c \geq 2$, even the normal approximation to the asymptotic is satisfactory. Note, however, that the asymptotic distribution (9) is itself a poor approximation to the correct finite n distribution for $c = 6$ or $c = 7$, because of the overwhelming number of cells having a zero count, and the low probability of any cell containing more than one observation. In fact, the values $X^2 = 13982018$ and $X^2 = 85898786$ observed when $c = 6$ in Table 1 and $c = 7$ in Table 2 respectively are the smallest possible values that this statistic could take in Lotto 6/49, and hence the real p -value is 1 in both cases. In general, when the number n of draws is less than $\binom{N}{k}$ in a k/N lottery, the probability of X^2 taking its smallest value $x_{\min} = \binom{N}{k} - n$ is the same as the probability that these first n draws yield different outcomes, i.e.

$$\begin{aligned}
 P(X^2 = x_{\min}) &= \prod_{i=1}^n \left\{ 1 - (i-1) / \binom{N}{k} \right\} \\
 &\approx \exp \left\{ - \sum_{i=1}^n (i-1) / \binom{N}{k} \right\} = \exp \left\{ - \binom{n}{2} / \binom{N}{k} \right\},
 \end{aligned}$$

so for Canada's Lotto 6/49, for instance, $P(X^2 = 13982018) > \frac{1}{2}$ for $n \leq 4403$, i.e. of the order of 42 years at the rate of two draws per week. It is not surprising, therefore, that no two draws have yet had the same outcome in this lottery's 20-year history.

Although it may be seen through similar calculations that the asymptotic approximation to the distribution of X^2 slowly deteriorates as c increases, there is no reason to doubt its reliability for values $c = 2, 3, 4$ which are of most practical importance.

7. Conclusion

When using a large random sample of data to test the hypothesis of uniformity of subsets of $c = 1, \dots, k$ winning numbers in a lottery of the type k/N , Pearson's standard goodness-of-fit statistic X^2 has been shown to be approximately distributed, not in the usual χ^2 -form, but rather as a weighted linear combination of c independent χ^2 random variables. This distribution is easily and accurately approximated by a simple linear transform of a single χ^2 random variable, and the resulting test is more powerful than the alternative test of Joe (1993) for a realistic type of alternative.

On the basis of the data that are currently available, neither Pearson's statistic X^2 nor Joe's test (for which the p -values are larger than 0.66 for $c = 2$, for both the six- and seven-ball analyses) provide any serious ground for suspecting a lack of uniformity in the results of Canada's 20-year-old Lotto 6/49. There might be situations in which the assumptions behind the above analysis would not apply. For example, if the mechanism or the balls themselves are changed, suitable tests for equiprobability based on X^2 using the known changepoint could be developed, but the power reported here would no longer be valid. Also, in using X^2 or the alternative statistic of Joe (1993) on lottery data, we should bear in mind that aggregate statistics such as these are not designed to detect the (remote) possibility of serial dependence between successive draws. Nevertheless, the tests developed here are particularly well adapted to testing the randomness of QuickPick selections, which often run in the millions, even for a given draw.

Acknowledgements

This research was funded in part by the Natural Sciences and Engineering Research Council of Canada and by the 'Fonds pour la formation de chercheurs et l'aide à la recherche du Gouvernement du Québec'.

Appendix A

This appendix shows how to derive the asymptotic distribution (9) for Pearson's statistic (7) in the general case $1 \leq c \leq k$. As in the case $c = 2$ treated in Section 2, the proof hinges on the computation of the eigenvalues of the covariance matrix Σ (which does not depend on n , as explained earlier). This matrix is

$$\Sigma = \text{cov}\left(\frac{O - E}{\sqrt{n}}\right),$$

where O and E denote the vectors of observed and expected counts for all possible elements $s \in P_c$, listed in some order that plays no role in what follows. For this reason, entries of Σ are simply indexed by elements of P_c , as was done in Section 2.

Let s and t be two elements of P_c , and let A_s and A_t denote the events that a given draw includes the balls in s and t respectively. Let also $d = |s \cap t| = |t \cap s| \leq c$. The (s, t) -entry of Σ is then given by

$$\begin{aligned} \sigma(s, t) &= P(A_s \cap A_t) - P(A_s)P(A_t) \\ &= \frac{\binom{N-c-d}{k-c-d}}{\binom{N}{k}} - \left\{ \frac{\binom{N-c}{k-c}}{\binom{N}{k}} \right\}^2 \equiv r_d, \end{aligned} \quad (11)$$

with the understanding that $\binom{x}{y} = 0$ unless $0 \leq y \leq x$, and $\binom{0}{0} = 1$ as usual.

Now introduce square matrices $Q_0, \dots, Q_c$ each of size $\binom{N}{c}$ by setting

$$Q_d(s, t) = 1(|s \cap t| = c - d),$$

for all $s, t \in P_c$, i.e. for arbitrary subsets s and t of $\{1, \dots, N\}$ of size c . We may then write

$$\Sigma = \sum_{d=0}^c r_d Q_d.$$

As mentioned in Section 2, the Q_d s are the adjacency matrices of the so-called Johnson association scheme, and hence they have the same eigenspace decomposition (Yamamoto *et al.*, 1965). For this specific scheme, Delsarte (1973) showed that if ν_l is defined as in expression (10) for $0 \leq l \leq c$ then there is a set of ν_l linearly independent vectors of length $\binom{N}{c}$ which are eigenvectors of $Q_0, \dots, Q_c$ simultaneously.

Delsarte further proved that these ν_l eigenvectors correspond to the same eigenvalue (of multiplicity ν_l) of Q_d , given by

$$\lambda_{d,l} = \sum_{j=0}^d (-1)^j \binom{l}{j} \binom{c-l}{d-j} \binom{N-c-l}{d-j}.$$

Consequently, Σ has at most $c+1$ different eigenvalues, given by

$$\kappa_l = \sum_{d=0}^c r_d \lambda_{d,l}, \quad 0 \leq l \leq c. \quad (12)$$

As for the case $c=2$, $\kappa_0=0$ and the formula can be simplified considerably for $1 \leq l \leq c$, but first it may be worth pointing out that $\lambda_{d,l} = E(d, l)$ with

$$E(d, x) = \sum_{j=0}^d (-1)^j \binom{x}{j} \binom{c-x}{d-j} \binom{N-c-x}{d-j},$$

which is the so-called Eberlein polynomial (Eberlein, 1964).

To reduce formula (12), first note that on changing the order of summation

$$\begin{aligned} \sum_{d=0}^c \lambda_{d,l} &= \sum_{d=0}^c \sum_{j=0}^d (-1)^j \binom{l}{j} \binom{c-l}{d-j} \binom{N-c-l}{d-j} \\ &= \sum_{j=0}^c (-1)^j \binom{l}{j} \sum_{d=j}^c \binom{c-l}{d-j} \binom{N-c-l}{d-j} \\ &= \sum_{j=0}^c (-1)^j \binom{l}{j} \sum_{x=0}^{c-l} \binom{c-l}{x} \binom{N-c-l}{x}. \end{aligned}$$

The sum over j effectively runs only to l , and the inner sum over $x = d-j$ is equal to $\binom{N-2l}{c-l}$ by Vandermonde's convolution formula (see, for example, Constantine (1987), page 6, or Riordan (1979), page 8). Consequently,

$$\sum_{d=0}^c \lambda_{d,l} = \binom{N-2l}{c-l} \sum_{j=0}^l (-1)^j \binom{l}{j} = \binom{N-2l}{c-l} (1-1)^l,$$

which vanishes except when $l=0$, in which case the sum is $\binom{N}{c}$. As a result,

$$\kappa_0 = \binom{N}{k} \sum_{d=0}^c \binom{c}{d} \binom{N-c}{d} \binom{N-c-d}{k-c-d} - \binom{N}{c} \binom{N-c}{k-c}^2.$$

But

$$\binom{N-c}{d} \binom{N-c-d}{k-c-d} = \binom{k-c}{d} \binom{N-c}{N-k},$$

so, through another application of Vandermonde's identity,

$$\begin{aligned} \kappa_0 &= \binom{N}{k} \binom{N-c}{N-k} \sum_{d=0}^c \binom{c}{d} \binom{k-c}{k-c-d} - \binom{N}{c} \binom{N-c}{k-c}^2 \\ &= \binom{N}{k} \binom{N-c}{N-k} \binom{k}{k-c} - \binom{N}{c} \binom{N-c}{k-c}^2 = 0, \end{aligned}$$

as can be readily checked.

Taking $1 \leq l \leq c$ from now on, we have

$$\kappa_l = \left\{ 1 / \binom{N}{k} \right\} \sum_{d=0}^c \binom{N-c-d}{k-c-d} \sum_{j=0}^d (-1)^j \binom{l}{j} \binom{c-l}{d-j} \binom{N-c-l}{d-j}.$$

Changing the order of summation and setting $x = d - j$ as above, we find that

$$\kappa_l = \left\{ 1 / \binom{N}{k} \right\} \sum_{j=0}^l \sum_{x=0}^{c-l} (-1)^j \binom{l}{j} \binom{c-l}{x} \binom{N-c-l}{x} \binom{N-c-x-j}{N-k}.$$

Reversing the order of summation once more, and using the fact that the inner sum is

$$\sum_{j=0}^l (-1)^j \binom{l}{j} \binom{N-c-x-j}{N-k} = \binom{N-c-l-x}{N-k-l}$$

(see, for example, Riordan (1979), formula (5a), page 8) we obtain

$$\kappa_l = \left\{ 1 / \binom{N}{k} \right\} \sum_{x=0}^{c-l} \binom{c-l}{x} \binom{N-c-l}{x} \binom{N-c-l-x}{N-k-l}.$$

At this point, the fact that

$$\binom{N-c-l}{x} \binom{N-c-l-x}{N-k-l} = \binom{N-c-l}{N-k-l} \binom{k-c}{x}$$

may then be exploited to write

$$\kappa_l = \left\{ 1 / \binom{N}{k} \right\} \binom{N-c-l}{N-k-l} \sum_{x=0}^{c-l} \binom{c-l}{x} \binom{k-c}{x},$$

which immediately reduces to

$$\kappa_l = \binom{k-l}{k-c} \binom{N-c-l}{k-c} / \binom{N}{k} \quad (13)$$

through a final application of Vandermonde's identity.

Finally, the weights $w_l = \kappa_l / e_c$, $1 \leq l \leq c$, may be deduced easily from formulae (8) and (13).

Remark 1. As already noted in the case $c = 2$, the eigenvector corresponding to κ_0 is the column vector $\mathbf{1}$ whose $\binom{N}{c}$ entries are equal to 1. The vector $\mathbf{1}$ is an eigenvector of all the Q_d s and hence of Σ . To see this, observe that, for fixed $s \in P_c$ and $0 \leq d \leq c$, we have

$$Q_d \mathbf{1}(s) = \sum_{t \in P_c} 1(|s \cap t| = c-d) = \binom{c}{d} \binom{N-c}{d} = \lambda_{d,0}, \quad (14)$$

so that $\Sigma \mathbf{1} = k_0 \mathbf{1}$ with

$$\kappa_0 = \sum_{d=0}^c r_d \binom{c}{d} \binom{N-c}{d}.$$

Since $\mathbf{1}' \Sigma \mathbf{1}$ is the variance of $\mathbf{1}'(O - E)\mathbf{1} = 0$, the fact that $\kappa_0 = 0$ does not come as a surprise.

References

- Bellhouse, D. R. (1982) Fair is fair: new rules for Canadian lotteries. *Can. Publ. Poly.*, **8**, 311–320.
- Constantine, G. M. (1987) *Combinatorial Theory and Statistical Design*. New York: Wiley.
- Delsarte, P. (1973) An algebraic approach to the association schemes of coding theory. *Philips Res. Rep.*, suppl. 10.
- Eberlein, P. J. (1964) A two parameter test matrix. *Math. Comput.*, **18**, 296–298.
- Fellegi, I. P. (1980) Approximate tests of independence and goodness of fit based on stratified multistage samples. *J. Am. Statist. Ass.*, **75**, 261–268.
- Holt, D., Scott, A. J. and Ewings, P. D. (1980) Chi-squared tests with survey data. *J. R. Statist. Soc. A*, **143**, 303–320.
- Imhof, J. P. (1961) Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.
- Joe, H. (1993) Tests of uniformity for sets of lotto numbers. *Statist. Probab. Lett.*, **16**, 181–188.
- Rao, J. N. K. and Scott, A. J. (1987) On simple adjustments to chi-square tests with sample survey data. *Ann. Statist.*, **15**, 385–397.
- Riordan, J. (1979) *Combinatorial Identities*. Huntington: Krieger.
- Searle, S. R. (1998) Winning probabilities of Lotto in the United States. *Chance*, **11**, 20–24.
- Stern, H. and Cover, T. M. (1989) Maximum entropy and the lottery. *J. Am. Statist. Ass.*, **84**, 980–985.
- Yamamoto, S., Fujii, Y. and Hamada, N. (1965) Composition of some series of association algebras. *J. Sci. Hiroshima Univ. A–I*, **29**, 181–215.