Table 2. Relative Frequency Estimates of p_{ij} Based on Series History

		NU	OU	OSU	CU	ISU	MU	KU	KSU	Win[a]
	NU	1	.428	.914	.734	.831	.591	.763	.849	.091
	OU	.572	1	.819	.802	.918	.677	.698	.824	.134
	OSU	.086	.181	1	.500	.617	.429	.500	.721	7×10^{-4}
$\mathbf{P}_{all}$ =	CU	.266	.198	.500	1	.733	.349	.594	.727	.003
	ISU	.169	.082	.383	.267	1	.360	.449	.639	1×10^{-4}
	MU	.409	.323	.571	.651	.640	1	.521	.736	.012
	KU	.237	.302	.500	.406	.551	.479	1	.703	.003
	KSU	.151	.176	.279	.273	.361	.264	.297	1	6×10^{-5}
	Lose[b]	3×10^{-5}	2×10^{-5}	.011	.006	.044	.003	.011	.121	

Pr (same order as 1988 finish) = 3.13×10^{-6} P_{pp} = .0038 ≈ 1/262

[a]Probability that a team will win all conference games.
[b]Probability that a team will lose all conference games.

however, lists the original score with Oklahoma as winner. A similar situation occurs with Kansas State and Kansas in 1980. Iowa State's guide lists a 1895 6-0 win over Missouri, while Missouri's guide starts their series records against Iowa State in 1896. Another interesting conflict concerns the 1907 Nebraska–Iowa State game. Nebraska claims a 10-9 win while Iowa State claims a 13-10 win. In this case each team gives a footnote indicating that the outcome of the game was disputed.

[*Received May 1989. Revised August 1990.*]

REFERENCE

Edgington, E. S. (1987), *Randomization Tests* (2nd ed.), New York: Marcel Dekker.

Weldon's Dice Data Revisited

ADRIENNE W. KEMP and C. DAVID KEMP*

This note examines the role of Weldon's classical dice data in illustrating the problems of model selection. New, more realistic models are discussed; simple Poisson trials models, with at most two unknown parameters, are fitted. An application to process-oriented inspection sampling for a multiple channel production process is suggested.

KEY WORDS: Constrained parameters; Multiple channel production process; Poisson trials; Unfair dice.

1. INTRODUCTION

Weldon's data on 26,306 throws of 12 dice are classic material for teaching Pearson's χ^2 goodness-of-fit test, both without and with an unknown parameter [see, for example, Fisher (1950, pp. 63–65 and 84–85), Yule and Kendall (1947, chap. 22)]. Recent treatments are Lloyd (1984, pp. 358–360) and Rayner and Best (1989, pp. 12–13 and 135–137).

The problems inherent in model formulation provide a wider role for these data. The customary replacement model for fair dice is equally unbalanced dice—we suggest that this is unrealistic. After describing the history of the data in Section 2, we put forward more general models for unfair dice in Section 3; these include the Poisson trials model. A parallel between these models and process-oriented inspection sampling for a multiple channel production process is drawn in Section 4. Section 5 introduces two forms of parameter constraint that simplify the Poisson trials model. Fits to the dice data are discussed in Section 6.

*Adrienne W. Kemp is Honorary Senior Lecturer and C. David Kemp is Honorary Professor, both in the Department of Mathematical and Computational Sciences, University of St. Andrews, St. Andrews, Scotland, KY16 9SS.

2. THE WELDON DICE DATA

Weldon obtained the data from 26,306 throws of 12 dice. He had two aims. One aim was to examine empirically the possibility of a finite number of causes leading to a continuous asymmetrical frequency curve via an asymmetric binomial. The other aim was to pinpoint the need for a valid criterion "to judge whether the differences between a series of group frequencies and a theoretical law, taken as a whole, were or were not more than might be attributed to the chance fluctuations of random sampling."

In a letter to Galton written on February 2, 1894, Weldon enclosed the data for all 26,306 throws and asked

Table 1. 12 Dice Thrown 26,306 Times: A Throw of 5 or 6 Reckoned a Success

Number of successes	Observed (1)	Expected binomial, $p = 1/3$ (2)	(1) − (2)	Expected EU, $\hat{p} = .338$	Expected ODD, $p = 1/3$, $\hat{P} = .385$	Expected LLO, $\hat{q} = .995$, $\hat{c} = .524$
0	185	203	−18	187	187	187
1	1,149	1,217	−68	1,146	1,145	1,147
2	3,265	3,345	−80	3,215	3,215	3,217
3	5,475	5,576	−101	5,465	5,467	5,466
4	6,114	6,273	−159	6,269	6,272	6,270
5	5,194	5,018	176	5,115	5,116	5,114
6	3,067	2,927	140	3,043	3,041	3,041
7	1,331	1,254	77	1,330	1,328	1,329
8	403	392	11	424	423	423
9	105	87	18	96	96	96
10	14	13	1	15	15	15
11	4	1	3	1	1	1
12	0	0	0	0	0	0
Total	26,306	26,306		26,306	26,306	26,306
Pearson's χ^2		35.5		8.2	8.3	8.2
df		10		9	9	8

his opinion on their validity; Galton's reply seems to be lost. Weldon put the data in a form where the recorded event was the number of dice showing 5 or 6 points, that is, the expected chance of "success" for each die was 1/3 (see Table 1). Of the 26,306 throws a subset of 7,006 (our Table 2) was made by a clerk at University College whom Weldon thought very reliable and accurate. He was worried because Karl Pearson had said that the deviation from the theoretically most probable result for these 7,006 throws was so great as to make the records "intrinsically incredible." Yule's notes, taken at Pearson's lectures in the autumn of 1894, give an account of Pearson's reasoning (he had not yet developed his χ^2 goodness-of-fit test). Edgeworth, however, wrote to Weldon on February 7, 1894, saying, "The tests which I have applied . . . do not yield a result which excites much suspicion." Details of Pearson's and Edgeworth's arguments are in E. S. Pearson's (1965) paper on the history of statistics from 1890 to 1894.

Two further subsets of the data are in Yule and Kendall (1947, chaps. 6 and 10). The data for 4,096 throws of 12 dice, with 6 points reckoned a success ($p = 1/6$), are reproduced in Table 3. The results for 4,096 tosses, with 4, 5, or 6 reckoned a success ($p = 1/2$), are given in Table 4.

In 1900 Karl Pearson published his seminal paper on the χ^2 goodness-of-fit statistic; the originality and importance of this paper were discussed in Plackett (1983). Weldon's complete data set is the basis of illustrations I and II of K. Pearson (1900). In illustration I, a binomial model with $n = 12$, $p = 1/3$ is shown to give a clear run of negative discrepancies followed by a run of positive discrepancies (see our Table 1) and a χ^2 value of 43.9. Pearson's inference is that "the odds are 62,499 to 1 against such a system of deviations on a random selection." In illustration II, the probability of a success is estimated as $p = .3377$ for each of the 12 dice. His recalculated value of χ^2 is 17.8, leading to the conclu-

Table 2. 12 Dice Thrown 7,006 Times: A Throw of 5 or 6 Reckoned a Success

Number of successes	Observed (1)	Expected binomial, $p = 1/3$ (2)	(1) − (2)	Expected EU, $\hat{p} = .337$	Expected ODD, $p = 1/3$, $\hat{P} = .373$	Expected LLO, $\hat{q} = .997$, $\hat{c} = .516$
0	45	54	−9	51	51	51
1	327	324	3	310	310	310
2	886	891	−5	864	864	864
3	1,475	1,485	−10	1,462	1,463	1,462
4	1,571	1,671	−100	1,670	1,670	1,670
5	1,404	1,336	68	1,356	1,356	1,356
6	787	780	7	803	803	803
7	367	334	33	350	349	350
8	112	104	8	111	111	111
9	29	23	6	25	25	25
10	2	4	−2	4	4	4
11	1	0	1	0	0	0
12	0	0	0	0	0	0
Total	7,006	7,006		7,006	7,006	7,006
Pearson's χ^2		16.4		12.0	12.1	12.0
df		10		9	9	8

Table 3. 12 Dice Thrown 4,096 Times: A Throw of 6 Reckoned a Success

Number of successes	Observed (1)	Expected binomial, $p = 1/6$ (2)	(1) − (2)	Expected EU, $\hat{p} = .167$	Expected ODD, $p = 1/6$, $\hat{P} = .166$	Expected LLO, $\hat{q} = .999$, $\hat{c} = .201$
0	447	459	−12	460	460	460
1	1,145	1,103	42	1,103	1,103	1,103
2	1,181	1,213	−32	1,213	1,213	1,213
3	796	808	−12	808	808	808
4	380	364	16	363	363	363
5	115	116	−1	116	116	116
6	24	27	−3	27	27	27
7	8	5	3	5	5	5
8+	0	1	−1	1	1	1
Total	4,096	4,096		4,096	4,096	4,096
Pearson's χ^2		5.5		5.5	5.5	5.5
df		7		6	6	5

sion that "the odds are now only 8 to 1 against a system of deviations as improbable as or more improbable than this one."

This has become standard treatment for the data, for example, by Yule and Kendall (1947), Fisher (1950), and Lloyd (1984). Pearson made his calculations without grouping small frequencies and with the expected frequencies rounded to the nearest integer. Different modes of grouping and degrees of rounding have led to the publication of various χ^2 values; our own calculations give $\chi^2_{[10]} = 35.5$ when $p = 1/3$ and $\chi^2_{[9]} = 8.2$ when p is estimated by maximum likelihood. The inferences are unaltered.

The University College clerk's 7,006 throws with 5 or 6 reckoned a success do not seem to have become teaching material. The chi-squared goodness-of-fit test with $n = 12$, $p = 1/3$ gives $\chi^2_{[10]} = 16.4$ (see Table 2); the dice appear to be fair. Similarly, fitting a binomial model with $n = 12$, $p = 1/6$ to the data in Table 3 gives $\chi^2_{[7]} = 5.5$ and, consequently, no cause to doubt the null hypothesis of fair dice (Yule and Kendall 1947, exercise 22.1).

However, when Yule and Kendall (1947, example 22.7) fitted a binomial model with $n = 12$, $p = 1/2$ to the data in Table 4, they obtained $\chi^2_{[12]} = 33.8$ (their assignment of degrees of freedom may be questioned—our own calculations give $\chi^2_{[10]} = 33.5$). There is little doubt that these data are inconsistent with a null hypothesis of fair dice. Yule and Kendall commented on the striking pattern of the signs of the discrepancies and referred back to their remark that if the discrepancies "are all negative in the cells farthest removed from the mean, [then] the standard deviation may show an almost impossible divergence from expectation." They suggested that supplementary tests should be applied, including those for the significance of the mean and standard deviation. If, unlike Yule and Kendall, we fit a binomial model with $n = 12$ and p estimated from the data, then $\chi^2_{[9]} = 7.5$ and the pattern of signs among the discrepancies vanishes; the data are consistent with this new null hypothesis.

Rayner and Best (1989) took the analysis of the complete data set a stage further. They ingeniously partitioned Pearson's χ^2 into orthonormal components that are asymptotically standard normal and showed that the first two such components provide tests for a shift in mean and a shift in variance. Their test for a mean shift

Table 4. 12 Dice Thrown 4,096 Times: A Throw of 4, 5, or 6 Reckoned a Success

Number of successes	Observed (1)	Expected binomial, $p = 1/2$ (2)	(1) − (2)	Expected EU, $\hat{p} = .512$	Expected ODD, $p = 1/2$, $\hat{P} = .641$	Expected LLO, $\hat{q} = .915$, $\hat{c} = 1.71$
0	0	1	−1	1	1	1
1	7	12	−5	9	9	9
2	60	66	−6	55	54	51
3	198	220	−22	191	189	185
4	430	495	−65	450	448	447
5	731	792	−61	754	755	759
6	948	924	24	921	924	933
7	847	792	55	827	829	835
8	536	495	41	541	542	540
9	257	220	37	252	251	246
10	71	66	5	79	78	75
11	11	12	−1	15	15	14
12	0	1	−1	1	1	1
Total	4,096	4,096		4,096	4,096	4,096
Pearson's χ^2		33.5		7.5	7.1	6.7
df		10		9	9	8

is the usual standardized mean test; they got $z = (\bar{x} - n/3)/(2n/9k)^{1/2} = 5.20$, with $n = 12$, $p = 1/3$, and $k = 26{,}306$, strongly indicating a mean shift. Their test for a variance shift is related though not, in general, identical to a standardized index of dispersion test. Unfortunately, the relevant normalized Krawtchouk polynomials are incorrectly stated on page 136 of Rayner and Best (1989), as is their $h_2(x; p)$. Our own calculations give

$$z = \sum [(x - np)^2 - np(1 - p) + (2p - 1)(x - np)]/[p(1 - p)\{2kn(n - 1)\}^{1/2}] = .75,$$

not 1.54; no variance shift is indicated.

Rayner and Best examined only the complete data set. The results of similar analyses applied to the three subsets of the data are as follows. For the University College clerk's 7,006 throws in Table 2, the z values for the standardized mean and Rayner–Best variance shift tests are 2.07 and 1.32, respectively; there is some evidence of a shift in the mean but not in the variance. For the data in Table 3, the two z values are $-.02$ and .30, respectively; the inference of a fair die is unaltered. For Table 4, the z values are 5.13 and $-.79$, respectively, with the inference that the data are inconsistent with a null hypothesis of fair dice, a shift in the mean but not in the variance being indicated. (Fisher's index of dispersion test also indicates no marked variance shift for any of the four data sets.)

3. MORE REALISTIC MODELS

Although the equally unbalanced dice hypothesis may give a satisfactory fit when the fair dice hypothesis fails to do so, the assumption that all 12 dice are identically unbalanced seems inherently implausible. More realistic hypotheses for unfair dice are now examined; we do not claim to exhaust the possibilities.

Consider k throws of n dice where p_{ij} is the probability of success for die number i on throw j. Then each toss of each die constitutes a Bernoulli trial with probability p_{ij}. Assume that the outcomes of each toss of each die are all independent.

For a binomial model, we require that $p_{ij} = p$ for all i, j. For each throw the number of successes X has the probability generating function (pgf) $(1 - p + ps)^n$, with mean np and variance $np(1 - p)$.

The so-called Poisson trials model arises when $p_{ij} = p_i$ ($j = 1, \ldots, k$; $i = 1, \ldots, n$) and not all p_i are equal (Poisson 1837). The pgf for the number of successes per throw is now $\Pi_i(1 - p_i + p_i s)$, the mean is $\Sigma_i p_i = n\bar{p}$ say, and the variance is $n\bar{p}(1 - \bar{p}) - \Sigma_i(p_i - \bar{p})^2 = n\bar{p}(1 - \bar{p}) - n\sigma_W^2$ say, which is obviously less than $n\bar{p}(1 - \bar{p})$. Here σ_W^2 is the within-throw variance of the p_i. The term Poisson binomial is sometimes used for this model, a usage that can be confused with the Poisson binomial distribution with pgf $\exp\{\lambda(1 - p + ps)^n - \lambda s\}$.

Lexis (1877) put forward an alternative extension of the binomial model, with $p_{ij} = p_j$ ($i = 1, \ldots, n$; $j = 1, \ldots, k$). The pgf for the number of successes per throw is now $\Sigma_j(1 - p_j + p_j s)^n/k$, with mean $\Sigma_j np_j/k = n\bar{p}$ say, and variance $n\bar{p}(1 - \bar{p}) + n(n - 1)\Sigma_j(p_j - \bar{p})^2)/k = n\bar{p}(1 - \bar{p}) + n(n - 1)\sigma_B^2$ say, which exceeds that of a binomial $(n, \bar{p})$ variable. Here σ_B^2 is the between-throws variance of the p_j. Stuart and Ord (1987, pp. 165 and 171) pointed out that this second generalization is a special case of cluster sampling and, in addition, that it can be regarded as a mixed binomial model.

Consider now the situation where p_{ij} vary both within and between throws. Here the pgf for the number of successes per throw is $\Sigma_j\Pi_i(1 - p_{ij} + p_{ij}s)/k$, the mean is $\Sigma_i\Sigma_j p_{ij}/k = n\bar{p}$ say, and the variance is

$$n\bar{p}(1 - \bar{p}) + n^2 \sum_j (p_j - \bar{p})^2/k - \sum_i\sum_j (p_{ij} - \bar{p})^2/k.$$

This will often be very similar to the variance for the Lexian model (see Aitken 1945, pp. 53–54).

Aitken gave a detailed presentation of the formulas for all four models; he called the fourth model Coolidge's extension of the Lexian scheme. The relationships between Bernoulli and Poisson trials having the same overall expected number of successes, with the somewhat surprising result that the variance is *less* for Poisson trials, are the subject of a lucid article by Nedelman and Wallenius (1986). The application of models based on Poisson trials to weapon defense systems has been studied recently by Thomas and Taub (1982) and Kemp (1987).

How do these models relate to the dice data?

If, for example, some of the dice were unfair at the start of a run of throws and the probability of success for each die remained constant throughout the throws, then Poisson trials would result. This could well be an appropriate model for the data in Table 4, where the sample mean is 6.139 and the sample variance of 2.931 is less than the theoretical variance of 3 for the fair dice model.

If the probability of success changed during the experiment in such a way that the dice were all equally unbalanced at any stage of the experiment, then a mixed binomial (Lexian) model would hold. This would occur if the dice were initially all identical and if during the course of the throws the n dice changed in such a manner that they remained identical after each of the k throws—an unlikely scenario.

However, if the probabilities of success were to alter during the course of the throws without remaining identical for all 12 dice, then the Coolidge model would hold. This could well be appropriate for the complete data set (Table 1), where the run of throws is an aggregation of several runs, the sample mean is 4.052, and the calculated variance is 2.698 compared with the theoretical variance of 2.667 for the fair dice model.

The difficulty with the Coolidge model is the impossibility of estimating the excessively large number of parameters from the data. Ways of overcoming a similar problem for the Poisson trials model are studied in Section 5.

4. MULTIPLE CHANNEL PRODUCTION PROCESSES

A single-sampling plan for attributes inspection involves taking a random sample of specified size n from each lot submitted for inspection. The items are then subjected to a prescribed test, and the number that do not conform to the product specification is counted. If this number is greater than a specified acceptance number, then the entire lot is rejected; if it is less than or equal to the specified acceptance number, the lot is accepted. If all items have the same probability p of rejection, then the variable of interest is binomial (n, p). A record of lots rejected/accepted enables the production process to be monitored.

Vardeman (1986) played devil's advocate in his recent overview of the role of inspection in modern quality control by asking whether inspection sampling is not "just spending resources playing dice?" Conversely, throwing dice can be regarded as simulating inspection sampling, especially if interest is shifted from product acceptance to process monitoring. We recall that Weldon's interest in dice tossing with $p \neq 1/2$ arose as an empirical investigation, that is, simulation study, in his (abortive) search for an asymmetrical continuous limiting distribution.

Many modern technological processes produce a fixed number, n, of similar items simultaneously. For example, in plastic intrusion molding a single mold may have n compartments, each of which produces a small plastic component such as a button or a bottle top. The obvious and easy method to check whether the mold has become unacceptably worn is to inspect its output of n items at regular points in time and count how many satisfactory items out of n the mold has produced.

A binomial model, where the probability p_i that a satisfactory item has been produced is the same for all n compartments, may no longer be appropriate, however. For instance, the items coming from the outermost compartments may be more prone to distortion during ejection from the mold. A null hypothesis that the mold is not subject to wear corresponds to the assumption that the not-all-equal probabilities p_i $(i = 1, 2, \ldots, n)$ remain constant from one inspection occasion to another, that is, to a Poisson trials model.

5. MODELS WITH FEWER PARAMETERS

The drawback to the nonbinomial models in Section 3 is the difficulty of estimating their many parameters. For the Lexian model this is customarily overcome by assuming that the p_j $(j = 1, 2, \ldots, k)$ are realizations from a distribution with very few parameters, for example, the beta distribution. We have seen, however, that this type of model is inappropriate in the present context. A Poisson trials model is more promising. Ways of making it more tractable by constraining the parameters p_i $(i = 1, 2, \ldots, n)$ are now examined.

A very simple special case of Poisson trials arises if we assume that all of the dice are fair except one; we call this the "one-dud-die" hypothesis. The pgf is now

$$(1 - P + Ps)(1 - p + ps)^{n-1},$$

where p is the probability of success for each of the $n - 1$ fair dice and P is the probability of success for the unfair die. Expanding and identifying the coefficient of s^x gives

$$\Pr(X = x) = (1 - P)\binom{n-1}{x}(1 - p)^{n-1-x}p^x + P\binom{n-1}{x-1}(1 - p)^{n-x}p^{x-1}.$$

Alternatively, differentiating the pgf with respect to s and equating coefficients of s^x gives the recurrence relation

$$(x + 1)\Pr(X = x + 1) = \left(\frac{P}{1 - P}(1 - x) + \frac{p}{1 - p}(n - 1 - x)\right)\Pr(X = x) + \frac{Pp}{(1 - P)(1 - p)}(n - x + 1)\Pr(X = x - 1).$$

Sequential computation of the probabilities for any (P, p) combination is straightforward using either of these results. The mean and variance are

$$\mu = P + (n - 1)p,$$
$$\sigma^2 = P(1 - P) + (n - 1)p(1 - p).$$

When $p > P \geq .5$ or $.5 \geq P > p$ the variance is greater than for the fair die model; for $P > p \geq .5$ or $.5 \geq p \geq P$ it is less.

Suppose now that there is a spectrum of unfairness among the dice. A log-linear odds assumption concerning the p_i gives

$$\ln[p_i/(1 - p_i)] = \ln c + (i - 1)\ln q, \quad i = 1, 2, \ldots, n,$$

that is,

$$p_i = cq^{i-1}/(1 + cq^{i-1}), \quad i = 1, 2, \ldots, n.$$

This is essentially the parametric model introduced by Cox (1958) to give a reasonable approximation to relationships likely to occur in practice and to have mathematical tractability.

The pgf is now

$$\prod_{i=0}^{n-1}(1 + cq^is)/(1 + cq^i),$$

with just two unknown parameters. Note that we assume that there is no record regarding which dice produced successes (which compartments produce unsatisfactory items), and so the ordering of the dice is immaterial. The restriction $q < 1$ can be imposed without loss of generality, since for any pair of parameters c, q, with $q > 1$, there exists a corresponding pair c', q', with $q' < 1$, where $c' = cq^{n-1}$, $q' = q^{-1}$, giving the identical distribution.

This distribution also arises as the exact outcome of a stationary stochastic process for the dichotomy between parasites on hosts with and without open wounds resulting from previous parasite attacks (Kemp and Newton 1990). It is not the same, however, as in Kemp (1987), where a weapon defense system model led to Poisson trials with a deterministic log-linear relationship between the probabilities of the form

$$\ln p_i = \ln C + (i - 1) \ln Q, \qquad i = 1, 2, \ldots, n.$$

An assumption of log-linear relationship between the probabilities is questionable for the dice data and for multiple channel production outputs, however, as it involves an unduly wide spread among the probabilities at one end of their observed range. Suppose, for example, that $n = 12$, $C = .1$, $Q = 1.2$; then the p_i are .100, .120, .144, .173, .207, .249, .299, .358, .430, .516, .619, .743. For the log-linear odds model with $n = 12$, $c = .1$, $q = 1.2$, the probabilities are .091, .107, .126, .147, .172, .199, .230, .264, .301, .340, .382, .426.

The pgf for the log-linear odds model can be restated in terms of basic hypergeometric functions (see Kemp and Newton 1990), whence the application of Heine's theorem [see, for example, Slater (1966, p. 92)] enables the pgf to be expanded as a series in s, giving

$$\Pr(X = x) = \frac{(1 - q^n)(1 - q^{n-1}) \cdots (1 - q^{n-x+1})}{(1 - q)(1 - q^2) \cdots (1 - q^x)} \times q^{x(x-1)/2} c^x \Big/ \prod_{i=1}^{n} (1 + q^{i-1} c),$$

($x = 0, 1, \ldots, n$; $q < 1$). Numerical calculation of the probabilities is very straightforward, as

$$\Pr(X = 0) = \prod_{i=0}^{n-1} (1 + cq^i)^{-1}$$

and

$$\Pr(X = x) = (q^{x-1} - q^n)c/(1 - q^x) \Pr(X = x - 1), \qquad x = 1, 2, \ldots, n.$$

The cumulants for the convolution of Bernoulli distributions are the sums of the corresponding individual cumulants; hence

$$\mu = \sum_{i=0}^{n-1} \{cq^i/(1 + cq^i)\} = n - \sum_{i=0}^{n-1} \{1/(1 + cq^i)\},$$

$$\sigma^2 = \sum_{i=0}^{n-1} \{cq^i/(1 + cq^i)^2\} = n - \mu - \sum_{i=0}^{n-1} \{1/(1 + cq^i)^2\}.$$

The variance can be either greater than for the fair dice model (as in Tables 1 and 2) or less (as in Tables 3 and 4).

As $q \to 1$, then $p_i \to c/(1 + c)$ ($i = 1, \ldots, n$), and the distribution becomes binomial (n, $c/(1 + c)$). The equally-balanced (EU) model is, therefore, a special case of the log-linear-odds (LLO) model; the fair dice model is a special case of both the EU model and the one-dud-die (ODD) model.

6. RESULTS AND DISCUSSION

Consider estimation of the unknown parameters in the Poisson trials models. For the EU model the moment estimator of p is $\bar{x}/n$ and, in addition, is the MLE. For the ODD model the moment estimator of P is $\bar{x} - (n - 1)p$, but is not the MLE, the calculation of which requires iteration. For the LLO model the moment estimators of c and q are again not the same as the MLE's; here both forms of estimation require iteration. To compare fits of the models MLE was chosen. In the case of the ODD and LLO models this was carried out using the appropriate recurrence relationship for the probabilities to compute log-likelihoods and a computer optimization algorithm based on the IMSL minimization procedure ZXMWD to search the log-likehood surface for the MLE(s); for details of ZXMWD see IMSL Library (1984).

For the data in Table 1 we recall that there is strong evidence against the hypothesis of fair dice ($p = 1/3$ for all dice). The EU model gives $\hat{p} = .338$ for all dice, whereas the ODD model estimates the unfair die probability P as .385. Under the LLO model with $\hat{q} = .995$ and $\hat{c} = .524$, the estimated values of p_i ($i = 1, \ldots, 12$) are .332, .333, .334, .335, .336, .337, .338, .339, .340, .342, .343, .344. The fits for the EU, LLO, and ODD models are all very alike, as are the χ^2 values, and the data can be said to be consistent with any one of them. If *only* these three alternative models are considered, then the choice between them hinges on one's attitude toward their realism. As stated in Section 3, however, a Coolidge model may be more appropriate, but this cannot be fitted without making some kind of assumption about the way in which the dice varied during the 26,306 throws.

We found only weak evidence for rejection of the null hypothesis of fair dice ($p = 1/3$) for the data in Table 2 (these are the 7,006 tosses that worried Weldon). The EU estimate of p is .337 (for all dice); the ODD estimate of P is .373. The LLO estimate of q is .997, which is close to unity; together with $\hat{c} = .516$ it gives the following nearly equal estimated values for p_i: .333, .334, .334, .335, .336, .336, .337, .338, .338, .339, .340, .340. The three χ^2 values are very close, and the fits are only slightly better than for the fair dice model. Their similarity is remarkable, despite their different p_i values.

Retention of a null hypothesis of fair dice is indicated for Table 3. The EU estimate is $\hat{p} = .167$, compared with $p = 1/6$ for the fair dice model, whereas the ODD estimate of P is .166. The estimated values q and c for the LLO model are $\hat{q} = .999$ and $\hat{c} = .201$, giving .166, .166, .166, .166, .166, .167, .167, .167, .167, .167, .167, .167 for the $\hat{p}_i$. The four χ^2 values are all very alike, as are the four fits.

We recall that Yule and Kendall (1947) commented that the pattern of signs for the fit for the fair dice model in Table 4 suggests underdispersion. The observed variance is indeed less than for the fair die model. The Rayner–Best tests indicate a mean shift, but not a variance shift. Here $\hat{p} = .512$ for EU (compared with $p = 1/2$) and $\hat{P} = .641$ for ODD. The LLO model with $\hat{q} = .915$, $\hat{c} = 1.709$ gives a wide spectrum of estimated values for p_i, namely .391, .413, .434, .456, .479, .501, .523, .545, .567, .589, .610, .631. The χ^2 values for the three models are very similar and are much lower than for the fair dice model.

For each data set the more realistic LLO model gives fits that are at least as good as those for the EU model. This is only to be expected, as the EU model is the special case of the LLO model where $q = 1$. Perhaps more surprisingly, the ODD model (which is *not* a special case of the LLO model) also did as well. Clearly, when the fair dice model is untenable, these more realistic models are worth consideration. On the basis of χ^2 goodness-of-fit tests (or indeed from consideration of log-likelihoods), however, the EU model cannot be rejected in favor of the LLO or ODD models for these data. Other models are of course possible, but model selection for such data sets clearly cannot be conducted on the basis of goodness-of-fit tests alone, even with data sets as large as these.

Finally, we recall that the EU, ODD, and LLO models are all Poisson trials models. The closeness of the results when fitting these models suggests that other Poisson trials models would give very similar results and that, unless one is interested in the parameters per se, the extra effort involved in estimation for more complicated Poisson trials models is unlikely to be worthwhile.

[*Received January 1988. Revised August 1990.*]

REFERENCES

Aitken, A. C. (1945), *Statistical Mathematics* (4th ed.), Edinburgh: Oliver & Boyd.

Cox, D. R. (1958), "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society,* Ser. B, 20, 215–232.

Fisher, R. A. (1950), *Statistical Methods for Research Workers* (11th ed.), Edinburgh: Oliver & Boyd.

IMSL Library (1984), *Math/Library. FORTRAN Subroutines for Mathematical Analysis, Version 1.0,* Houston, TX: Author.

Kemp, A. W. (1987), "A Poissonian Binomial Model With Constrained Parameters," *Naval Research Logistics Quarterly,* 34, 853–858.

Kemp, A. W., and Newton, J. (1990), "Certain State-Dependent Processes for Dichotomised Parasite Populations," *Journal of Applied Probability,* 27, 251–258.

Lexis, W. (1877), *Zur Theorie der Massenerscheinungen in der Menschlichen Gesellschaft,* Freiburg: Wagner.

Lloyd, E. H. (1984), "Chi-squared Tests of Goodness of Fit and of Independence and Homogeneity," in *Handbook of Applicable Mathematics, Vol. VI Statistics, Part A,* ed. E. H. Lloyd, Chichester, U.K.: John Wiley, chap. 7.

Nedelman, J., and Wallenius, T. (1986), "Bernoulli Trials, Poisson Trials, Surprising Variances, and Jensen's Inequality," *The American Statistician,* 40, 286–289.

Pearson, E. S. (1965), "Some Incidents in the Early History of Biometry and Statistics 1890–94," *Biometrika,* 52, 3–18.

Pearson, K. (1900), "On the Criterion that a Given System of Deviations From the Probable in the Case of a Correlated System of Variables Is Such That it Can be Reasonably Supposed to Have Arisen From Random Sampling," *Philosophical Magazine, 5th Series,* 50, 157–175.

Plackett, R. L. (1983), "Karl Pearson and the Chi-Squared Test," *International Statistical Review,* 51, 59–72.

Poisson, S. D. (1837), *Recherches sur la Probabilité des Jugements,* Paris: Bachelier.

Rayner, J. C. W., and Best, D. J. (1989), *Smooth Tests of Goodness of Fit,* New York: Oxford University Press.

Slater, L. J. (1966), *Generalized Hypergeometric Functions,* Cambridge, U.K.: Cambridge University Press.

Stuart, A. and Ord, J. K. (1987), *Kendall's Advanced Theory of Statistics, Vol. 1* (5th ed.), London: Charles Griffin.

Thomas, M. A., and Taub, A. E. (1982), "Calculating Binomial Probabilities When the Trial Probabilities Are Unequal," *Journal of Statistical Computation and Simulation,* 14, 125–131.

Vardeman, S. B. (1986), "The Legitimate Role of Inspection in Modern SQC," *The American Statistician,* 40, 325–328.

Yule, G. U., and Kendall, M. G. (1947), *An Introduction to the Theory of Statistics* (13th ed.), London: Charles Griffin.