Mathematical Statistics
Stockholm University

A note on “shaved dice” inference

Rolf Sundberg

Research Report 2016:12

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2016:12**,
http://www.math.su.se

A note on "shaved dice" inference

Rolf Sundberg[*]

June 2016

Abstract

Two dice are rolled repeatedly, but only their sum is registered. Have the two dice been "shaved", so two of the six sides appear more frequently? Pavlides & Perlman (2010) discuss this somewhat complicated type of situation through curved exponential families. Here we contrast their approach by regarding data as incomplete data from a simple exponential family. The latter, supplementary approach is in some respects simpler, it provides additional insight about likelihood equation and Fisher information, it opens up for the EM algorithm, and it elucidates the information content in ancillary statistics.

Key words: aggregated cells, ancillarity, curved exponential families, EM algorithm, Fisher information, incomplete data model, ML estimation, multinomial model

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@math.su.se Websites: http://www.su.se/profiles/rolfs and http://staff.math.su.se/rolfs/

1 Introduction

In an entertaining and instructive article, Pavlides & Perlman (2010) consider the following (artificial) inference problem. A pair of dice used at a casino are possibly "shaved", both identically, such that two mutually opposite sides of the die have a higher probability than the other four sides. A statistical complication is that the two dice are rolled together and the statistics collected "unfortunately" do not represent the individual dice outcomes, but only the sum of the two outcomes. Pavlides and Perlman show how this complication leads away from a simple binomial situation to a curved multinomial exponential family, with a scalar parameter but a minimal sufficient statistic of much higher dimension (4 or 5 for shaved ordinary dice). One consequence, typical for curved families, is that the likelihood equation does not have an explicit solution. Pavlides and Perlman demonstrate how the Fisher information can be calculated and used to judge the sample size needed for a desired precision in the parameter, in particular in comparison with individual dice data. They also show and illustrate how these results much depend on what pair of faces is assumed to possibly have been shaved, including how the dice were labelled.

The purpose of this note is to point out that another instructive way to look at such data is as incomplete data from an exponential family of individual die data. The results are of course not in conflict with the results of the curved approach, but they may be considered simpler in some sense, and they give different insights in the results. In particular they open up for the EM algorithm, including control of its speed of convergence, and for a more elucidating analysis of ancillarity.

2 Incomplete data approach

With observed data regarded as incomplete data from an exponential family with scalar canonical statistic t and canonical parameter θ, the likelihood equation can be written

$$\mathrm{E}[t \mid \mathrm{data}] = \mathrm{E}[t], \tag{1}$$

where the left hand side is the conditional expected value of t, given the observed incomplete data (Sundberg 1974). Both sides are functions of the parameter, which need not be the canonical θ, but the equation holds for any choice of parameterization. The observed information for the canonical parameter θ is

$$J(\theta;\, \mathrm{data}) = \mathrm{Var}[t] - \mathrm{Var}[t \mid \mathrm{data}],$$

and the corresponding Fisher information is the expected value of J,

$$I(\theta) = \mathrm{Var}[t] - \mathrm{E}[\mathrm{Var}[t \mid \mathrm{data}]] = \mathrm{Var}[\mathrm{E}[t \mid \mathrm{data}]].$$

In another parameterization, by $\psi = g(\theta)$, we have $I(\psi) = \mathrm{Var}[\mathrm{E}[t \mid \mathrm{data}]]/g'(\psi)^2$. In both cases, $\mathrm{Var}[\mathrm{E}[t \mid \mathrm{data}]]$ is expressed in terms of the adequate parameter. For complete data, $\mathrm{E}[t|\mathrm{data}] = t$, so $\mathrm{Var}[t \mid \mathrm{data}] = 0$ and $\mathrm{Var}[\mathrm{E}[t \mid \mathrm{data}]]$ simplifies to $\mathrm{Var}[t]$. It follows that for any parameterization, the relative Fisher information, as compared with complete data, is

$$I_{rel} = \mathrm{Var}[\mathrm{E}[t \mid \mathrm{data}]]/\mathrm{Var}[t], \tag{2}$$

where both numerator and denominator are functions of the parameter. The last part of theory we need is that $1 - I_{rel}$ is not only the expected relative loss of information due to incompleteness, but also the rate of convergence of the EM algorithm, which updates the parameter by solving Eq. (1) with the current parameter value on the left hand side and the new one on the right hand side (Sundberg 1976, Dempster et al. 1977).

More specifically, in the present case we can take $t = \overline{x} = \sum x_i/n$, where $x_i = 0$, $\frac{1}{2}$ or 1, is the proportion of dice in the ith of n rolls of the pair of dice showing any of the two larger faces (the shaved ones). Then $E[t]$ is the probability to get one of the larger faces when rolling a single die, which we can take as parameter itself (mean value parameterization) or express as a function of some other convenient parameter. Denoting the observed sum in the ith roll by y_i, the left hand side of the likelihood equation (1) can be written

$$\mathrm{E}[t \mid \mathrm{data}] = \sum_{i=1}^{n} \mathrm{E}[x_i \mid y_i]/n = \sum_{j=2}^{12} \mathrm{E}[x \mid y = j]\, f_j, \tag{3}$$

where $j = 2, \ldots, 12$ are the possible sum outcomes and f_j is the observed relative frequency of outcome j. Thus we need formulas for these $\mathrm{E}[x \mid y = j]$, expressed as functions of the parameter. For some j-values they will be immediat§e, equal to 0, $\frac{1}{2}$ or 1, whereas for other j-values we have to apply the definition of conditional probability to the expression

$$\mathrm{E}[x \mid y = j] = \tfrac{1}{2}\,\mathrm{Pr}(x = \tfrac{1}{2}|y = j) + \mathrm{Pr}(x = 1|y = j).$$

When we have this list of conditional mean values for all j (see example below), the computer easily calculates (3), which we need on the left hand side of (1). Replacing the f_j in (3) by the corresponding theoretical probabilities (known functions of the parameter, already used in the conditional probabilities), we would get $E[t]$, again.

Only slightly more complicated is to instruct the computer to calculate the observed and Fisher information quantities. We just need expressions for $\mathrm{Var}[x \mid y = j]$ instead of $\mathrm{E}[x \mid y = j]$ in (3). The Fisher information is obtained from the observed information by again replacing the f_j by the corresponding theoretical probabilities, and I_{rel} in (2) by dividing by $\mathrm{Var}[t]$. The procedure is exemplified in more detail in the next section. Finally, we could follow Pavlides & Perlman (2010, Fig. 3) and plot I_{rel} as function of the parameter.

3 Example

As an example of the calculations to be carried out in the “incomplete data” approach, we look at the first example of Pavlides & Perlman (2010), with two ordinary dice possibly shaved along one of the two opposite faces *1* and *6*. The other die variants considered by Pavlides & Perlman (2010) are treated analogously. We follow their notations, letting a be the probability for each of the faces *1* and *6*, and b the probability for each of the other faces, under the obvious constraint $2a + 4b = 1$. We need only consider the case of a single roll of the two dice ($n = 1$). By symmetry we could reduce the number of registered outcomes y by merging 2 and 12, 3 and 11 etc, but we need not do so for the procedure below.

First and foremost we need the conditional distribution of x, the average number of outcomes 1 or 6, given the sum y. For symmetry reasons we need only consider $y \leq 7$ here. For $y = 2$ and $y = 3$, x is nonrandom, $x = 1$ and $x = \frac{1}{2}$, respectively. For the other y-values we have a binary distribution, so we only need the probability for the non-zero outcome, and it will immediately yield the conditional means and variances for x. Here is a list of the conditional probabilities we need:

$$\begin{aligned}
\Pr(x = 1 \,|\, y = 2) &= a^2/a^2 = 1 \\
\Pr(x = \tfrac{1}{2} \,|\, y = 3) &= 2ab/2ab = 1 \\
\Pr(x = \tfrac{1}{2} \,|\, y = 4) &= 2ab/(2ab + b^2) \\
\Pr(x = \tfrac{1}{2} \,|\, y = 5) &= 2ab/(2ab + 2b^2) \\
\Pr(x = \tfrac{1}{2} \,|\, y = 6) &= 2ab/(2ab + 3b^2) \\
\Pr(x = 1 \,|\, y = 7) &= 2a^2/(2a^2 + 4b^2)
\end{aligned}$$

They are here expressed such that the denominator of the ratio form is the probability for the particular y-value, the same as in Pavlides & Perlman (2010, eqs (2.1)). The conditional variance is the non-zero x-value squared times the product of the conditional probability and its complement. So for example, $\mathrm{Var}[x \,|\, y = 2] = \mathrm{Var}[x \,|\, y = 3] = 0$, and

$$\mathrm{Var}[x \,|\, y = 4] = (\tfrac{1}{2})^2\,(2ab)\,(b^2)\,/(2ab + b^2)^2. \tag{4}$$

Its contribution to the expected $\mathrm{E}[\mathrm{Var}[x \,|\, y]]$ is obtained by multiplying by $\Pr(y = 4)$, that is by deleting one of the factors $(2ab+b^2)$ in the denominator of (4). Calculating $\mathrm{E}[\mathrm{Var}[x \,|\, y]]$ in this way yields the following formula:

$$1 - I_{rel} = \frac{b}{4}\left\{\frac{1}{2a+b} + \frac{1}{a+b} + \frac{3}{2a+3b} + \frac{4a}{a^2+2b^2}\right\} \tag{5}$$

This of course resembles the formula for $e_{1,6}(a)$ in Pavlides & Perlman (2010), since I_{rel} and $e_{1,6}(a)$ are necessarily identical. The terms on the right hand side tells how the incompleteness in the different y-values implies loss of information. For $a = b = 1/6$ the right hand side becomes $(1/24)\{2 + 3 + 3.6 + 8\} = 0.69$. The four terms correspond to $y = 4$ or 10, $y = 5$ or 9, $y = 6$ or 8, and $y = 7$, respectively. It is clear that the incompleteness in $y = 7$ contributes most to the loss of information. We return to this matter in Section 4.

As calculated already by Pavlides & Perlman (2010), $e_{1,6}(1/6) = 0.31$. This implies that if we use the EM algorithm to compute the MLE $\widehat{a}$ of a, assumed close to $a = 1/6$, we should expect the deviation from $\widehat{a}$ to decrease by a factor $1 - 0.31 = 0.69$ per iteration. In other words, one more decimal will be right per about 6 iterations ($0.69^6 = 0.11$).

4 Ancillarity

For precision estimation in inference with actual data we can choose between observed and expected information, or even better a conditional expected information, if there is an ancillary statistic (precision index) to condition on. Pavlides & Perlman (2010, Sec. 5.6) discuss the existence of alternative ancillary statistics and the choice between them for a non-conventionally labelled type of die. The incomplete data approach also elucidates ancillarity effects, and to show this, we continue on the example above. First, a division of the possible outcomes y in the sets of even and odd values is easily seen to be ancillary, in the sense of both frequencies being binomial with (parameter-free) probability $\frac{1}{2}$, and as we will see it is also a precision index. A further division of the even y-values in those < 7 and those > 7 is also distribution-constant in the same sense, but does not form a useful precision index. Minimal sufficiency, or symmetry, tells us to aggregate over $y = 7 - j$ and $y = 7 + j$.

Let even and odd be represented by the statistic u. First we note that $E(x|\, u)$ is independent of u. Hence the conditional mean value of the MLE is at least approximately independent of u, whereas $\mathrm{Var}(x|\, u)$ depends on u, so u is a proper precision index. The former property actually holds for any distribution-constant division of the set of y-values for any labelling of the dice. Formula (5) for $1 - I_{rel}$ quantifies u as a precision index. The first and third of the four terms correspond to $u =$ "even". From the numerical values when $a = 1/6$ it is seen that the by far largest information loss is when $y = 7$, which is in the set $u =$ "odd". Thus, if the odd y-values are less frequent than expected (i.e.$< 50\%$), we should expect a higher precision than expected on average over u, but a lower precision if the odd outcomes are $> 50\%$.

The situation is even more extreme in Sec. 5.6 of Pavlides & Perlman (2010), where faces *1* and *2* have the possibly higher probability a. The set u of y-values $\{2, 5, 9\}$ in their division S_1 has probability $1/4$, independent of a, and $\mathrm{Var}(x \mid u) = 0$. In other words, for outcomes in u there is no loss of information, so a high frequency of such outcomes is highly beneficial, and vice versa if the frequency is low, and this is quantified by conditioning on the ancillary S_1.

5 Discussion

The curved family analysis of Pavlides & Perlman (2010) and the incomplete data analysis above contribute in different ways to the understanding of the models. The incomplete data analysis seems to be simpler, since it only involves a one-dimensional sufficient statistic. The derivations of the likelihood equations and the Fisher information are perhaps not considerably simpler, but they appear to yield more understanding of the results than the curved family approach.

Situations in which the duality between curved exponential families and incomplete data from simple exponential families should be considered are not infrequent. More generally than in the example of Pavlides & Perlman (2010), they often appear in contingency tables when some cells are aggregated. In particular this is so in genetics, where the aggregation may correspond to one genotype dominating another genotype, thus controlling the observed phenotype. Two illustrations are the introductory example of Dempster et al. (1977), where four cells (categories) have the probabilities $\{(2+\pi)/4,\ (1-\pi)/4,\ (1-\pi)/4,\ \pi/4\}$ for some parameter π, $0 < \pi < 1$, and the classical example of blood-groups A, B and 0, modelled under Hardy–Weinberg equilibrium.

Under a relative degree of incompleteness as in the example above, the EM algorithm is a serious competitor to the Newton–Raphson algorithm for calculating the MLE, because EM converges reasonably fast, is easier to program, and has a lower risk for divergence since the likelihood increases for each iteration. Note that with a different design of the dice, the rate of convergence will be different, whereas the sample size n does not have an effect on the formula for the EM expected rate of convergence.

We have also seen that the incomplete data approach helps to understand the effects of conditioning on ancillary statistics. It should be kept in mind, however, that if we really want to make a precise inference about shaved dice, very large sample sizes are needed, and then with high probability the ancillary statistics will be close to their expected values. As a consequence, the conditional variance given an ancillary statistic is likely to be close to its expected value.

References

Dempster, A., Laird, N. & Rubin, D. (1977), ‘Maximum likelihood from incomplete data via the EM algorithm (with discussion)’, *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

Pavlides, M. & Perlman, M. (2010), ‘On estimating the face probabilities of shaved dice with partial data’, *The American Statistician* **64**(1), 37–45.

Sundberg, R. (1974), ‘Maximum likelihood theory for incomplete data from an exponential family’, *Scand. J. Statist.* **1**, 49–58.

Sundberg, R. (1976), ‘An iterative method for solution of the likelihood equations for incomplete data from exponential families’, *Comm. Statist. – Sim. Comp. B* **5**, 55–64.