Estimating Paired-Interaction Probabilities

S. S. Wagner
ESTIMATING PAIRED-INTERACTION PROBABILITIES

S.S. WAGNER

Yerkes Regional Primate Research Center, Emory University, Atlanta, GA

Suppose that the various members of some group of \( K \) individuals, called senders, are directing acts of some kind toward the various members of that or another group of \( K \) individuals, called receivers, and that, during some period of observation, they made a total of \( N \) such acts. There are then \( K^2 \) different sender-receiver combinations, and hence \( K^2 \) possible outcomes at each trial, or occurrence of a single act. The problem is to account for the observed distribution of the \( N \) acts among the \( K^2 \) possible outcomes. The individuals may be countries, people, monkeys, or whatever, the acts may be trade shipments, threats, groomings or whatever, and the receivers may be the same as or different from the senders; for the sake of concreteness, however, we shall discuss the problem in terms of monkeys and groomings, with the group of senders the same as the group of receivers.

Letting \( n_i \) and \( m_j \) denote the number of times monkey \( i \) groomed and was groomed respectively, and \( n_{ij} \) the number of times monkey \( i \) groomed monkey \( j \), we have

\[
\begin{align*}
n_i &= n_{i1} + \ldots + n_{iK}, \quad m_i = n_{1i} + \ldots + n_{Ki}, \quad n_1 + \ldots + n_K = m_1 + \ldots + m_K = N.
\end{align*}
\]

Such data are sometimes displayed in the form of a 'contingency table'

\[
\begin{array}{cccc}
& n_{11} & \ldots & n_{1K} \\
& n_{21} & \ldots & n_{2K} \\
& \vdots & \ddots & \vdots \\
& n_{K1} & \ldots & n_{KK} \\
m_1 & m_2 & \ldots & m_K \\
N
\end{array}
\]
whence the \( n_i \)'s and \( m_j \)'s may be called the row and column totals (or marginals) respectively.

If the groomings occur independently of one another they will have a multinomial distribution. That is, if at each trial there remains an unchanging probability \( p_{ij} \) that it will be monkey \( i \) grooming monkey \( j \), then the likelihood \( l \) of obtaining, in \( N \) trials, exactly the distribution observed will be given by

\[
\begin{align*}
\hat{l} &= c \cdot \frac{n_{11} \cdot n_{12} \cdots n_{1K} \cdot n_{K1} \cdots n_{KK}}{N!} \\
\end{align*}
\]

where \( c = \frac{1}{n_{11}! \cdot n_{12}! \cdots n_{KK}!} \).

We include the cases where some \( p_{ij} = 0 \) by agreeing that \( 0^0 = 1 \). The \( p_{ij} \)'s can be arranged in a table in the same way as the \( n_i \)'s.

In the case of an ordinary table (one with no \( p_{ij} \)'s known a priori to be zero), one may then test the null hypothesis that there is independence between sender and receiver, i.e., that the various ratios obtainable between \( p_{ij} \)'s in a given row are the same for all rows, which is the same as the hypothesis that \( p_{ij} = p_i \cdot q_j \) for all \( i \) and \( j \) — where the row total \( p_i = p_{i1} + \cdots + p_{iK} \) is the probability that monkey \( i \) will groom, and the column total \( q_j = p_{1j} + \cdots + p_{Kj} \) is the probability that monkey \( i \) will be groomed. This is called the independence-model, or hypothesis of independence. Under it, (1) becomes

\[
\begin{align*}
\hat{l} &= c \cdot p_i \cdots p_K \cdot q_i \cdots q_K \\
\end{align*}
\]

which, except for \( c \), no longer depends on the individual entries \( n_i \) but only upon the marginals \( n_i \) and \( m_j \). The foregoing is well-known and discussed in many textbooks of probability and statistics.

In the case of a non-ordinary table (one or more \( p_{ij} \)'s known a priori to be zero), one may test the null hypothesis of quasi-independence between sender and receiver, i.e., that the various ratios obtainable between non-a-priori-zero \( p_{ij} \)'s in a given row are the same for all rows wherever applicable, which is equivalent to the hypothesis that there exist non-negative parameters \( P_i \) and \( Q_j \) with \( P_1 + \cdots + P_K = Q_1 + \cdots + Q_K = 1 \) such that \( p_{ij} = \frac{P_i \cdot Q_j}{t} \) for every \( p_{ij} \) not a priori zero — where \( t = 1 - \sum_{(i,j) \not\in \Gamma} P_i \cdot Q_j = \sum_{(i,j) \not\in \Gamma} P_i \cdot Q_j \).
with the index set \( \Gamma \) defined by: \((i, j) \in \Gamma \) only if \( p_{ij} \) a priori zero. (Division by
the total, \( t \), is necessary to normalize the \( p_{ij} \), that is, to obtain \( \sum p_{ij} = 1 \).) Thus,
for example, if grooming is never self-directed (the zero-diagonal case), \( t \) will
equal \( 1 - P_1Q_1 - \ldots - P_KQ_K = \sum p_{ij} \), while if grooming cannot proceed down
the order-hierarchy (the upper-right-triangle-zero case), \( t \) will equal \( 1 - \sum_{i < j} p_{ij} \)

\[ \sum_{i > j} p_{ij} \]  This is called the quasi-independence-model, or hypothesis of
quasi-independence. The parameters \( P_i \) and \( Q_i \) may be called the ‘theoretical
tendencies’ of monkey \( i \) to groom and be groomed respectively; they are not
in general equal to the true tendencies (probabilities) \( p_i = \sum_j p_{ij} \) and \( q_i = \sum_j p_{ij} \)
respectively. Under this model, (1) becomes

\[
L = \prod_{i=1}^{n_1} p_i \cdot \prod_{j=1}^{m_K} q_j = \prod_{i=1}^{n_K} \frac{m_i}{n_i} \cdot \prod_{j=1}^{m_K} \frac{m_j}{m_K}
\]

which, as before, no longer depends upon the individual entries \( n_i \).

In order to test either the independence-model or the quasi-independence
model, one must have at hand some estimate of the parameters involved. The
estimates ordinarily sought are the ‘maximum likelihood’ estimates, that is,
in the first case, the \( p \)’s and \( q \)’s making \( L \) as large as possible, and, in the second
case, the \( P \)’s and \( Q \)’s making \( L \) as large as possible. Now it is well-known that
the unique solution to the first problem is given by

\[
p_i = \frac{n_i}{N} \quad \text{and} \quad q_i = \frac{m_i}{N}
\]

for all \( i \); for the second, however, no general solution is known.

In [1] we give all solutions for the zero-diagonal case, showing that for the
usual table of \( n_j \)’s, the solution is unique and may be obtained by finding the
value of \( \alpha \) for which \( K - 2 - R_1 - \ldots - R_K = 0 \), where

\[
R_i = \sqrt{\left( 1 - \frac{n_i + m_i}{\alpha} \right)^2 - \frac{4n_i m_i}{\alpha^2}} \quad \text{and setting}
\]

\[
P_i = \frac{1}{2} \left( 1 + \frac{n_i - m_i}{\alpha} - R_i \right) \quad \text{and} \quad Q_i = \frac{1}{2} \left( 1 + \frac{m_i - n_i}{\alpha} - R_i \right)
\]
for all $i$; in [2] we give all solutions for the upper-right-triangle-zero case, showing that, for the usual table, the solution is unique and may be obtained by setting

$$P_1 = \frac{n_1}{m_1}, \quad Q_k = \frac{m_k}{n_k}$$

$$P_2 = \frac{n_2(1-P_1)}{m_1 + m_2 - n_1}, \quad Q_{k-1} = \frac{m_{k-1}(1-Q_k)}{n_{k-1} + n_k - m_k}$$

$$P_3 = \frac{n_3(1-P_1 - P_2)}{m_1 + m_2 + m_3 - n_1 - n_2}, \quad Q_{k-2} = \frac{m_{k-2}(1-Q_{k-1} - Q_k)}{n_{k-2} + n_{k-1} + n_k - m_{k-1} - m_k}$$

etc.

A general procedure, which would include these two as well as other cases, is obviously to be desired.

Having the $P$'s and $Q$'s, one can find $t$, and hence the $p_{ij}$'s. The predicted or expected values for the table are then given by $E_{ij} = p_{ij} \cdot N$ and may be compared with the observed values by means of a $\chi^2$-test (df = $K^2 - 2K + I$ - number of cells in $I$) to see whether the disagreement is significant enough to cause rejection of the model.

SUMMARY

One of the assumptions underlying a standard model for the distribution of hits, threats, etc. among a group of monkeys is that each monkey $i$ has 'theoretical tendencies', $P(i)$ and $Q(i)$, to hit and be hit respectively. Various iterative procedures have been suggested for finding the maximum likelihood values of the parameters $P(i)$ and $Q(i)$, but none of them are known to be convergent, nor to yield a solution if convergent. This paper discusses solutions to two cases of that problem.

REFERENCES


Author's address: Stephen S. Wagner, Yerkes Primate Research Center, Emory University, Atlanta, GA 30322 (USA).