ESTIMATING THE PARAMETERS OF THE BETA-BINOMIAL DISTRIBUTION

Rand R. Wilcox
Center for the Study of Evaluation
University of California at Los Angeles

The work upon which this publication is based was performed pursuant to a grant [contract] with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

For certain types of tests for which an item sampling model applies and when the items are scored dichotomously, the beta-binomial distribution (BBD), also known as the negative hypergeometric or Polya distribution, may be useful when describing the marginal distribution of observed scores for a particular population of examinees (Keats and Lord, 1962; Lord, 1965; Lord and Novick, 1968, Chapter 23). Wilcox (1977) and Huynh (1976), for example, have used this distribution to describe a mastery test.

The BBD arises as follows: Consider the binomial probability function

$$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{1}$$

If we view p as arising from a beta distribution $g(p)$ with parameters r and s, the marginal distribution of x is

$$\begin{aligned} f(x) &= \int \binom{n}{x} p \, (1-p)^{n-x} g(p)dp \\ &= \binom{n}{x} \frac{B(r+x,\ n-x+s)}{B(r,\ s)} \end{aligned} \tag{2}$$

where B is the usual beta function. In mental test theory p is sometimes referred to as the percent correct true score of an examinee and $g(p)$ is the distribution of true scores over a population of examinees. Various properties and extensions of (2) are given by Ishii and Hayakawa (1960).

A fundamental problem when using the BBD is estimation of the parameters r and s. Two methods are typically employed -- an approximation to the maximum likelihood estimate (MLE) and the method of moments. Results reported by Shenton (1950) suggest that the two estimation procedures will yield nearly the same results when the number of examinees is

From Keats and Lord (1962) we see that the moment estimates may be written as

$$\hat{r}_1 = (-1+\alpha_{21}^{-1})\hat{\mu}$$

$$\hat{s}_1 = -\hat{r}_1+n/\alpha_{21}-n$$

where α_{21} is the KR21 reliability coefficient and $\hat{\mu}$ is the sample mean of the observed data.

MAXIMUM LIKELIHOOD ESTIMATES

Skellam (1948) describes an iterative procedure for obtaining maximum likelihood estimates of r and s using the digamma function. More recently, Griffiths (1973) has shown that the digamma function can be avoided. Instead, one determines the value of π and θ, say $\hat{\pi}$ and $\hat{\theta}$, so that

$$(3) \qquad \sum_{i=0}^{n-1} \frac{S_n-S_i}{\pi+i\theta} - \frac{S_{n-1-i}}{1-\pi+i\theta} = 0$$

$$(4) \qquad \sum_{i=1}^{n-1} i \left\{\frac{S_n-S_i}{\pi+i\theta} + \frac{S_{n-1-i}}{1-\pi+i\theta} - \frac{S_n}{1+i\theta}\right\} = 0$$

where $S_i = \sum_{x=0}^{i} f_x$ and where f_x is the number of examinees obtaining an observed x.

The estimates of r and s are

$$(5) \qquad \hat{r}_2 = \hat{\pi}\hat{\theta}^{-1}$$

$$(6) \qquad \hat{s}_2 = \hat{\theta}^{-1} - \hat{\pi}\hat{\theta}^{-1}$$

Solutions to (3) and (4) can be obtained iteratively using the Newton-Raphson method. The method of moments gives a good starting point.

$\pi = r(r+s)^{-1}$ and $\theta = (r+s)^{-1}$. The log likelihood for the BBD is given by

$$\ln L = c - S_n \sum_{i=1}^{n-1} \ln(1+i\theta) + \sum_{i=0}^{n-1} \{(S_n - S_i)\ln(\pi + i\theta) +$$

$$S_{n-1-i} \ln(1-\pi+i\theta)\}$$

where the constant c becomes zero after differentiation and so plays no role in the estimation of r and s, as will become apparent. Define the vector G as

$$G = \left(\frac{\partial \ln L}{\partial r}, \frac{\partial \ln L}{\partial s}\right)$$

and the 2x2 matrix Λ as

$$\Lambda = \begin{pmatrix} \dfrac{\partial^2 \ln L}{\partial^2 r} & \dfrac{\partial^2 \ln L}{\partial r\ \partial s} \\ \\ \dfrac{\partial^2 \ln L}{\partial s\ \partial r} & \dfrac{\partial^2 \ln L}{\partial^2 s} \end{pmatrix}$$

and consider

(7) $$(\hat{r}_3, \hat{s}_3) = (\hat{r}_1, \hat{s}_1) - G\Lambda^{-1}$$

The statistics $\hat{r}_3$ and $\hat{s}_3$ are simply the values obtained after one iteration using the Newton-Raphson procedure.

We verify that (7) is BAN (best asymptotically normal). As noted by Blischke (1964, p. 522) it is sufficient to verify that $\hat{r}_1$ and $\hat{s}_1$ are $o(\sqrt{k})$-consistent. That is, we must verify that $k^{\frac{1}{2}}(\hat{r}_1 - r)$ and $k^{\frac{1}{2}}(\hat{s}_1 - s)$ are bounded in probability uniformly in k. To show that $\hat{r}_1$ and $\hat{s}_1$ are $o(\sqrt{k})$-consistent we need only show that they are asymptotically normal (Ferguson, 1958). Applying a theorem given by Cramér (1946, p. 366) $\hat{r}_1$ and $\hat{s}_1$ are asymptotically normal and so (7) is a BAN estimate of r and s.

(10) $\psi(r) \doteq \ln(r-\frac{1}{2})$.

For $0 < r < 2$ we use

(11) $\psi(r) \doteq \psi(r+2)-r^{-1} - (r+1)^{-1}$

where ψ on the right-hand side of (11) is given by (10). Proceeding in a similar manner we have that

$$\Gamma''(r) = \Gamma(r)\psi'(r) + \Gamma'(r)\,\psi(r)$$

For ψ' we use the approximation

$$\psi'(r) \doteq (r-\tfrac{1}{2})^{-1},\ r \geq 2$$
$$\doteq \psi(r+2) + r^{-2} + (r+1)^{-2},\ 0 < r < 2.$$

The fifth estimation procedure is the jacknife (Quenouille, 1956) using one-at-a-time omission. The purpose of this estimator is to reduce the bias in the moment estimators $\hat{r}_1$ and $\hat{s}_1$. Miller (1974) gives a good review of this approach to estimation.

Let $\tilde{r}_i$ be the moment estimate of r with the ith observation removed. That is, $\tilde{r}_i$ is computed in the same manner as $\hat{r}_1$ using the observations $x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k$. The statistic $\tilde{s}_i$ is defined in an analogous fashion. The jacknife estimate of r is

$$\hat{r}_5 = k^{-1} \sum_{i=1}^{k} T_i$$

where

$$T_i = k\hat{r}_1 - (k-1)\tilde{r}_i$$

computer costs. In practice the Newton-Raphson procedure usually converges after only a few iterations, at least for a tolerance error of 10^{-4}. In all of our empirical studies we considered only situations where both r and s are less than 20. Experience with the BBD suggests that r and s will usually be in this range. (See, e.g., Shenton, 1950; Ishii and Hayakawa, 1960; Williams, 1975).[1]

Column one in Table 1 indicates the estimation procedure used. Thus, the results for estimator 2 are the results obtained when using $\hat{r}_2$ and $\hat{s}_2$. Consider, the example, the case r=s=.5. Table 1 indicates that $E(\hat{r}_2-r)^2=1.25$ and $E(\hat{s}_2-s)^2=.377$.

We decided to report the results for the values of r and s shown in Table 1 primarily because this demonstrates that the accuracy of all five estimation procedures might be poor when the number of examinees is small. In terms of comparing the various estimation techniques all combinations of r and s gave the same general results. In particular, we found that among the admissible estimates of r and s, the Newton-Raphson approximation to the MLE was most accurate of all. As can be seen, the improvement of $\hat{r}_2$ over the other four estimators, in terms of squared error loss, is often quite dramatic. A similar result was found for $\hat{s}_2$. Thus, to be conservative, it appears that an investigator should always compute an approximation to the MLE of r and s when the beta-binomial model is being used.

As previously indicated, experience suggests that $\hat{r}_1$ and $\hat{r}_2$ will usually give nearly the same results when the number of examinees is large. To find out whether this was the case for the situations considered here, we examined the 1,000 values of $\hat{r}_1$ and $\hat{s}_2$ that resulted from the 1,000 iterations in a

---

[1]Ishii and Hayakawa also report a case where both r and s are close to 100. Consideration of cases where r and s are large would not change the conclusions made here.

TABLE 1

Estimator	Probability of an admissible estimate of r	Probability of an admissible estimate of s	$E(\hat{r}-r)^2$	$E(\hat{s}-s)^2$
		(r, s) = (.5, .5)		
1	99.9	99.9	.45	.96
2	82.1	83.0	1.25	.377
3	96.3	87.8	.75	1.30
4	93.9	94.5	28.1	95.3
5	99.9	99.9	.46	.962
		(r, s) = (2, 2)		
1	97.4	97.4	1.5×10^4	1.5×10^4
2	77.5	79.1	3.5	3.4
3	90.1	92.8	23.6	24.5
4	66.5	64.8	2.2×10^6	2.2×10^6
5	97.4	97.4	1.5×10^4	1.5×10^4
		(r, s) = (5, 5)		
1	81.4	81.4	435.0	429.6
2	86.3	80.4	24.8	24.8
3	98.0	98.3	5.7×10^3	1.0×10^4
4	52.9	53.1	2.7×10^5	2.6×10^5
5	81.4	81.4	435.1	429.6
		(r, s) = (2, 9)		
1	80.3	80.3	9.2×10^{10}	1.5×10^{12}
2	86.1	79.8	3.33	65.7
3	79.9	80.3	247.8	5.3×10^3
4	52.8	51.7	1.4×10^{13}	2.3×10^{14}
5	80.3	80.3	9.2×10^{10}	1.5×10^{12}

Shenton, L. R. Maximum likelihood and the efficiency of the method of moments. Biometrika, 1950, 37, 111-116.

Skellam, J. G. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of the Royal Statistical Society, 1948, 10B, 257-261.

Wilcox, R. R. Estimating the likelihood of a false-positive or false-negative decision with a mastery test: an empirical Bayes approach. Journal of Educational Statistics 1977, 2, 289-307.

Williams, D. A. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics, 1975, 31, 949-952.

Wolfowitz, J. Asymptotic efficiency of the maximum likelihood estimator. Teoriia Veroiatnostei, 1965, 10, 267-281.