On Estimating Several Binomial N's

George Casella
Cornell University

and

William E. Strawderman
Rutgers University

BU-1171-M

August 1992

ON ESTIMATING SEVERAL BINOMIAL N'S

George Casella[1]
Cornell University

and

William E. Strawderman[2]
Rutgers University

Part of this research was performed at the Cornell University Workshop on Conditional Inference sponsored by the Army Mathematical Sciences Institute and The Statistics Center, Cornell University, Ithaca, New York, June 3-14, 1991.

Key Words and Phrases: Binomial distribution, estimation of sample size, admissibility.

AMS 1980 Subject Classification: 62F10, 62C20.

[1]Research supported by National Science Foundation Grant DMS 91-00839 and NSA Grant 90F-073.

[2]Research supported by National Science Foundation Grant DMS 90-23172.

Summary

We study the problem of estimating several Binomial sample sizes under the assumption that the true proportions are known and equal. The loss function is taken to be the sum of squared errors divided by the true sample sizes. The improved estimators are similar in form to the Clevenson-Zidek estimators in the problem of estimating several Poisson parameters. We also study the problem of estimating the expected value of several Binomials with equal but unknown proportions and unknown sample sizes. For this problem, we find an improved estimator if p is bounded above by some $p_0 < 1$. If p is unrestricted, the usual estimator is shown to be admissible.

1. Introduction

The problem of estimating the binomial n has received a fair bit of attention in the literature, perhaps due to the Fisher-Haldane disagreement (Fisher (1941-43); Haldane (1941-42)). More recently Olkin et al. (1981), Carroll and Lombard (1985) and Casella (1986) have all presented algorithms for n estimation. These algorithms are less concerned with decision theory, and more concerned with the instability of standard n estimators.

A decision-theoretic treatment of n estimation was given by Feldman and Fox (1968) who considered the single n case. Here we consider the problem of estimating several Binomial sample sizes. In particular, let $X_i \sim Bin(n_i, p)$, $i = 1, \ldots, k$. Suppose that the value of p is known and equal for each of the populations. We wish to estimate the vector $n = (n_1, n_2, \ldots, n_k)$ with loss

$$L(n, \delta) = \sum_{i=1}^{k} \frac{(\delta_i - n_i)^2}{n_i}. \tag{1.1}$$

This loss appears to be the most natural for this problem. Casella (1988) argues that estimating n is a scale parameter problem, and (1.1) is a natural scale parameter loss. Also, using (1.1), the usual estimator $X/p = (X_1/p, \ldots, X_n/p)$ has risk $k(1-p)/p$, constant in n_i. This is the estimator that we will improve upon, and we show that the estimator $\delta_a(X)(1 - \frac{a}{Z+k-1})\frac{X}{p}$ dominates $\frac{X}{p}$ for $0 < a < 2(k-1)(1-p)$ where $Z = \sum_{i=1}^{k} X_i$ and $k \geq 2$. This estimator is analogous to the Clevenson-Zidek (1975) estimator for estimating several Poisson means.

Johnson (1987) has considered the problem in a slightly different context. He allows the p_i's to be known but different and studies the loss $\sum_{n=1}^{k} k_i^2(\delta_i - n_i)^2$. His estimators are somewhat different than ours and his domination results are for $k \geq 3$.

We also study the related problem of estimating $np = (n_1 p, n_2 p, \ldots, n_k p) = EX$ in the case where p is unknown. Here the loss is

$$L(np, \delta) = \sum_{n=1}^{k} \frac{(\delta_i - n_i p)^2}{n_i p}. \tag{1.2}$$

We show that $X = (X_1 \ldots X_k)$ is inadmissible for its expected value provided that p is bounded above by a known constant $p_0 < 1$. The improved estimators are of the form

$(1 - \frac{a}{Z+k-1})X$ where $0 < a < 2(k-1)(1-p_0)$. The estimator X is shown to be admissible if p is unknown and unrestricted.

Although estimation of n_i assuming p is known is a problem that seemingly only has theoretical interest, our results can lead to usable n estimators in the unknown p case. In particular, Carroll and Lombard (1985) argue and provide convincing evidence) that integrating out p produces more stable n estimators. They use a variety of beta priors for p, and through simulation studies and data analysis demonstrate the superiority of their estimators. Therefore, in practice, an experimenter could consider an estimator of the form

$$\begin{aligned}\delta_a^*(X) &= \int_0^1 \delta_a(X)\pi(p)dp = \int_0^1 \left(1 - \frac{a}{k-1+Z}\right)\frac{X}{p}\pi(p)dp \\ &= \left(1 - \frac{a}{k-1+Z}\right)XE_\pi\frac{1}{p}\end{aligned}$$

where the prior π is a Carroll-Lombard prior.

2. Estimating Several Sample Sizes

This section is devoted to the main result of the paper.

THEOREM 2.1: *Let $X_i \sim Bin(n_i, p)$, $i = 1, \ldots, k$, where p is known $(0 < p < 1)$. The estimator $(\frac{X_1}{p}, \frac{X_2}{p}, \ldots, \frac{X_k}{p}) = \frac{X}{p}$ is inadmissible as an estimator of $(n_1, \ldots, n_k)$ with loss given by (1.1) if $k \geq 2$. A class of dominating estimators is given by*

$$\delta_a(X) = \left(1 - \frac{a}{k-1+Z}\right)\frac{X}{p} \tag{2.1}$$

where $0 < a < 2(k-1)(p-1)$ and $Z = \sum_{i=1}^k X_i$.

PROOF: We use the well-known fact that $X_i|Z$ has a hypergeometric distribution and, for $N = \sum_{i=1}^k n_i$, that

(i) $E(X_i|Z) = Zn_i/N$,

(ii) $Var(X_i|Z) = Z\frac{n_i}{N}(1 - \frac{n_i}{N})(\frac{N-Z}{N-1})$.

The difference in risks is given by

$$\Delta R = R(p, \delta_a) - R(p, \frac{X}{p}) \tag{2.2}$$

$$= E\Big\{E\Big[\Big(a^2\sum_{i=1}^{k}\frac{X_i^2}{p^2(k-1+Z)^2n_i} - 2a\sum_{i=1}^{k}\frac{X_i(X_i/p-n_i)}{p(k-1+Z)n_i}\Big)\Big|Z\Big]\Big\}$$

$$= E\Big\{E\Big[\Big(\frac{a}{p^2(k-1+Z)^2}\Big)\Big(\sum_{i=1}^{k}\frac{X_i^2}{n_i}(a-2(k-1+Z)) + \sum_{i=1}^{k}2(k-1+Z)X_ip\Big)\Big|Z\Big]\Big\}$$

$$= E\Big[\frac{a}{p^2(k-1+Z)^2}\Big(\sum_{i=1}^{k}\Big\{(a-2(k-1+Z))$$
$$\cdot\Big(\frac{Z^2}{N^2}n_i + \frac{Z}{N}\Big(1-\frac{n_i}{N}\Big)\Big(\frac{N-Z}{N-1}\Big)\Big)\Big\} + 2pZ(k-1+Z)\Big)\Big]$$

$$= E\Big[\frac{a}{p^2(k-1+Z)^2}\Big((a-2(k-1+Z))\Big(\frac{Z^2}{N} + \frac{Z(N-Z)(k-1)}{N(N-1)}\Big) + 2pZ(k-1+Z)\Big)\Big]$$

$$= \frac{a}{p^2N(N-1)}E\Big[\Big(\frac{Z}{k-1+Z}\Big)$$
$$\cdot\Big(\frac{(a-2(k-1+Z))((N-k)Z+N(k-1))}{k-1+Z} + 2pN(N-1)\Big)\Big].$$

Now use the fact that $(N-k)Z + N(k-1) \le N(k-1+Z)$ to obtain

$$\Delta R \le \frac{a}{p^2N(N-1)}E\Big(\frac{Z}{k-1+Z}\Big)\Big[aN - 2((N-k)Z + N(k-1) - N(N-1)p)\Big].$$

Next we use the fact (from the covariance inequality) that

$$E\frac{Z^2}{k-1+Z} \ge E\Big(\frac{Z}{k-1+Z}\Big)EZ = E\Big(\frac{Z}{k-1+Z}\Big)Np$$

to obtain

(2.3)

$$\Delta R \le \frac{a}{p^2N(N-1)}E\Big(\frac{Z}{k-1+Z}\Big)\Big[aN - 2((N-k)Np + N(k-1)) + 2N(N-1)p\Big]$$
$$= \frac{a}{p^2(N-1)}E\Big(\frac{Z}{k-1+Z}\Big)\Big[a - 2(k-1)(1-p)\Big].$$

Hence, $\Delta R < 0$ if $0 < a < 2(k-1)(1-p)$. Q.E.D.

REMARK 1: Essentially the same argument gives domination of estimators of the form $(1 - \frac{ar(Z)}{k-1+Z})X$ where $0 < a < 2(k-1)(1-p)$, $0 < r(Z) < 1$ and $r(Z)$ is nondecreasing in Z. The keys are that $r^2(Z) \le r(Z)$ and $\frac{r(Z)Z}{k-1+Z}$ is monotone nondecreasing. The final inequality for ΔR in the proof holds with $\frac{Z}{k-1+Z}$ replaced by $\frac{r(Z)Z}{k-1+Z}$.

3. Estimating the Vector of Expected Values

If instead of estimating $(n_1,\ldots,n_k)$ we consider the problem of estimating $(n_1p, n_2p,\ldots,n_kp) = (EX_1, EX_2,\ldots,EX_k)$ with loss (1.2), we can handle, to some degree at least, the situation where p is unknown. A natural estimator in this case would be $X = (X_1, X_2,\ldots,X_k)$. Theorem 3.1 shows that this estimator can be dominated by estimators of the form $(1 - \frac{a}{k-1+Z})X$ provided that $p \le p_0 < 1$.

THEOREM 3.1: *Let p be unknown and satisfy $0 < p \le p_0 < 1$. For the problem of estimating $(n_1p, n_2p,\ldots,n_kp)$ with loss equal to (1.2), the estimator $X = (X_1, X_2,\ldots,X_k)$ is beaten for $k \ge 2$ by $\delta_a^*(X) = (1 - \frac{ar(Z)}{k-1+Z})X$ provided*

(a) $0 < r(Z) \le 1$,

(b) $r(Z)$ *is nondecreasing,*

(c) $0 < a < 2(k-1)(1-p_0)$.

PROOF: This result follows directly from Theorem 2.1 and Remark 1 since

$$E\sum_{i=1}^{k}\frac{(\delta_{a,i}^*(X) - n_ip)^2}{n_ip} = pE\sum_{i=1}^{k}\frac{(\delta_{a,i}^*(X)/p - n_i)^2}{n_i} = pE\sum_{i=1}^{k}\frac{(\delta_{a,i}(X) - n_i)^2}{n_i}$$

where $\delta_{a,i}(X)$ is the ith component of the estimator $\delta_a(X)$ of Theorem 2.1. Hence $\delta_a^*(X)$ beats $\delta_0^*(X) = X$ for loss (1.2) iff $\delta_a(X)$ beats $\delta_0(X) = \frac{X}{p}$ for loss (1.1).

By Theorem 2.1, $\delta_a(X)$ dominates $\delta_0(X)$ provided $0 < a < 2(k-1)(1-p)$. The infimum of $2(k-1)(1-p)$ over the set $0 < p \le p_0 < 1$ is $2(k-1)(1-p_0)$. Q.E.D.

If p is unrestricted, the above theorem gives no dominating estimator. The following result shows that, in fact, no improvement is possible in this situation.

THEOREM 3.2: *Suppose p is unknown and unrestricted in the interval $[0,1]$. Then $X = (X_1,\ldots,X_k)$ is admissible for estimating $(n_1p, n_2p,\ldots,n_kp)$ for loss (1.2).*

PROOF: Let $X + g(X)$ be any competing estimator which is at least as good as X.

$$\begin{aligned} R(np, X + g(X)) =& E\left[\sum_{i=1}^{k}\frac{1}{n_ip}(X_i - n_ip)^2\right] \\ &+ E\left[\sum_{i=1}^{k}\frac{1}{n_ip}g_i^2(X)\right] + 2E\left[\sum_{i=1}^{k}(X_i - n_ip)g_i(X)\right]. \end{aligned}$$

If $p = 1$, then $X_i = n_i$ with probability one. Hence,

$$R(n1, X + g(X)) = \sum_{i=1}^{n} \frac{1}{n_i} g_i^2(n) > 0 = R(n1, X)$$

provided $g_i(n) > 0$ for any $n = (n_1, \ldots, n_k)$. Hence $g_i(X) \equiv 0$. Q.E.D.

Bibliography

Carroll, R. J. and Lombard, F. (1985). "A note on N estimators for the binomial distribution." *Journal of the American Statistical Association* **80**, 423-426.

Casella, George (1986). "Stabilizing binomial n estimators." *Journal of the American Statistical Association* **81**, 172-175.

Feldman, Dorian and Fox, Martin (1968). "Estimation of the parameter n in the binomial distribution." *Journal of the American Statistical Association* **63**, 150-158.

Fisher, R. A. (1941-43). "The negative binomial distribution." *Annals of Eugenics* **11**, 182-187.

Haldane, J. B. S. (1941-42). "The fitting of binomial distributions." *Annals of Eugenics* **11**, 179-181.

Johnson, R. W. (1987). "Simultaneous estimation of binomial N's." *Sankhya A* **49**, 264-267.

Olkin, I., Petkau, A. J., and Zidek, J. U. (1981). "A comparison of n estimators for the binomial distribution." *Journal of the American Statistical Association.* **76**, 639-642.