Jour. Ind. Soc. Ag. Statistics
48(2), 1996 : 185–193

On Sequential Estimation of Genetic Parameters and their Functions

K. Dutta and V. Goswami
Sambalpur University, Sambalpur-768019
(Received : July, 1993)

SUMMARY

In this paper the concept of sequential estimation of gene, genotype or phenotype proportions of a population has been introduced. The problem of sequential estimation of rare gene, genotype or phenotype proportions when the forces which influence the gene frequencies are present or absent have been discussed.

Key words : Gene, Genotype proportion, Phenotype proportion, Hardy-Weinberg law, Bernoulli population, Multivariate Bernoulli population, Sampling plan, Multinomial sampling plan, Inverse multinomial sampling plan, Average Sampling Number (ASN), Efficient estimator.

1. Introduction

Genes (alleles in a population) appear in pairs. We denote (for two alleles at a locus) dominant alleles by capital letters say 'A' and recessive alleles by small letters say 'a'. Let p and q represent the relative frequencies of alleles A and a respectively in the gene population. According to Hardy-Weinberg law, the zygotic combinations predicted in a randomly mating population may be represented by p^2: 2pq: q^2 proportions for AA,Aa and aa genotypes respectively [Refer Gardner [3] and Narain [4]]. This holds when the factors which influence the gene frequencies are absent but this is certainly not true for the present generation of living beings with critical changes of environmental effects. Therefore, we have discussed the method of estimation of rare genes proportions under two situations. In literature, the methods of estimation, discussed are based on fixed sample size (assuming the Hardy Weinberg law) even for estimating a rare trait which may not appear in a considerable large sized sample. Hence, in that case the sample may not contain any information about the rare trait to estimate its population proportion.

In this paper, we have introduced the sequential estimation procedure under two situations i.e. when the gene population is not directly influenced by the

forces like 'selection' and 'mutation' etc. and secondly when the population is influenced by these forces.

2. *Background*

The estimation of population proportions using multinomial and inverse multinomial sampling plans have been discussed in Dutta and Goswami [2]. We state here some relevent results which are necessary for the sequential estimation of gene proportions.

Consider a multivariate random variable $Y = (e_1, e_2, ..., e_{r+1})$

where $e_i = 0$ or 1 for i= 1, 2, .. (r+1) such that $\sum_{i=1}^{r+1} e_i = 1$

and $P(Y = y \text{ with } e_j = 1) = P_j > 0$ for $j = 1, 2, 3, ... (r+1)$

Let $y_1, y_2, y_3, ...$ be a sequence of observations on a random variable Y. We stop at taking N observations with the help of a stopping rule where N may be a random variable (it is constant for multinomial sampling plan).

Then $x = \sum_{i=1}^{N} y_i = (x_1, x_2, ..., x_{r+1})$

where x_i is the number of $e_i = 1$ (i = 1, 2,..., (r+1)) in the sample.

Suppose, our stopping rule is "continue taking observations up to when N = n (a pre-determined number)".

Hence, $\sum_{i=1}^{r+1} x_i = n$

and we get the multinomial sampling plan (MSP) with probability of reaching a boundary point as

$$P_m(X) = \left[n! / \prod_{i=1}^{r+1} x_i! \right] \prod_{t=1}^{r+1} P_t^{x_t} \qquad ...(2.1)$$

It is Binomial Sampling Plan (BSP) for r=1.

Suppose, we have a stopping rule as "continue taking observations in a sequence until a pre-determined number of observations fall into a given class which is rarest one". Without loss of generality, let us assume that we take observations just until $X_{r+1} = C$ have been observed. It generates the Inverse

Multinomial Sampling Plan (IMSP) with probability of reaching a boundary point as

$$P_{IM}(X) = \frac{\left(C + \sum_{i=1}^{r} x_i - 1\right)!}{(C-1)\prod_{i=1}^{r} x_i!} P_{r+1}^{C} \cdot \prod_{t=1}^{r} P_t^{x} t \qquad ...(2.2)$$

It reduces to Inverse Binomial Sampling Plan (IBSP) for r = 1.

The unique unbiased estimator of $p = (p_1, p_2, \ldots p_r)$ and the variance and co-variance expressions of the estimators under different sampling plans i.e. (i) MSP, (ii) IMSP, (iii) BSP and (iv) IBSP are stated below.

(i) Unbiased estimators of $p = (p_1, p_2 \ldots p_r)$ and their variances and co-variances in MSP

$$\hat{P}_i = (X_i/n),\ i = 1, 2, \ldots\ldots r \qquad ...(2.3)$$

$$V(\hat{p}_i) = p_i(1-p_i)/n,\ i = 1, 2, \ldots. r \qquad ...(2.4)$$

$$Cov(\hat{p}_i, \hat{p}_j) = -p_i p_s/n,\ i \neq j = 1, 2, \ldots r \qquad ...(2.5)$$

(ii) Unbiased estimators of $p = (p_1, p_2, \ldots, p_r)$ and their variances and co-variances in IMSP

$$\hat{P}_i = \left[x_i/(C-1+\sum_{i=1}^{r} x_i) \right],\ i = 1, 2, \ldots, r \qquad ...(2.6)$$

$$\hat{P}_{r+1} = (C-1)/(C-1+\sum_{i=1}^{r} x_i),\ \text{for } C > 1 \qquad ...(2.7)$$

$$V(\hat{P}_i) = Cp_i^2 f(C+2) + p_i p_{r+1} f(C+1) - p_i^2 \qquad ...(2.8)$$

$$i = 1, 2, \ldots, r$$

$$V(\hat{P}_{r+1}) = (C-1) P_{r+1}^2 f(C) - P_{r+1}^2 \qquad ...(2.9)$$

$$Cov(\hat{p}_i, \hat{p}_k) = C p_i p_k f(C+2) - p_i p_k \qquad ...(2.10)$$

$$i \neq k = 1, 2, \ldots, r$$

$$Cov(\hat{p}_i, \hat{p}_{r+1}) = (C-1) p_{r+1} p_i f(C+1) - p_i p_{r+1} \qquad ...(2.11)$$

$$i = 1, 2, \ldots, r$$

where $$f(C) = (C-2)! \sum_{j=0}^{\infty} \frac{j!\ (1-p_{r+1})^j}{(C-1+j)!} \qquad ...(2.12)$$

The unbiased estimators of p_1 and p_2 and their variances can be easily obtained from the expressions given in (i) and (ii) for BSP and IBSP substituting r=1.

We have mentioned here the unique unbiased estimators of quadratic functions of proportions like $p_i^2 + \beta p_i$ for i = 1, 2, ... (r+1) and $p_i p_j$ for $i \neq j = 1, 2,...r$ under different sampling plans. For r = 1, the quadratic functions $p_2^2 + \beta p_2$ takes the form of genotype proportions for different values of β i.e. (i) p_2^2 when $\beta = 0$ (ii) $p_2 (1-p_2)$ when $\beta = 1$ (iii) $(1-p_2)^2 - 1$ when $\beta = -2$ and the proportion of phenotype appearance of dominate gene as $-(p_2^2 + \beta p_2)$ when $\beta = -2$.

The variance expressions of unbiased estimators of $p_2^2 + \beta p_2$ in BSP and IBSP have been discussed in Dutta [1] and the variance expressions for estimators in MSP and IMPS can be derived with the help of treatments available there.

(iii) Unbiased estimators of $p_i^2 + \beta p_i$, i = 1, 2,... (r+1) and $p_i p_j$, i = j = 1, 2, ... (r + 1) under MSP

$$\hat{p}_i^2 + \beta \hat{p}_i = \frac{x_i (x_i - 1)}{n (n-1)} + \beta \frac{x_i}{n}, \quad i = 1, 2, \ldots (r+1) \qquad ...(2.13)$$

$$\hat{p}_i \hat{p}_j = \frac{x_i x_j}{n (n-1)}, \quad i \neq j = 1, 2, \ldots (r+1) \qquad ...(2.14)$$

(iv) Unbiased estimators of $p_i^2 + \beta p_i$, i = 1, 2,... (r+1), $p_i p_j$, i = j = 1, 2, ...r. and $p_i p_{r+1}$, i = 1, 2,... r under IMSP

$$\hat{p}_i^2 + \beta p_i = \frac{x_i (x_i - 1)}{(C-1+\sum_{i=1}^{r} x_i)(C-2+\sum_{i=1}^{r} x_i)} + \beta \frac{x_i}{(C-1+\sum_{i=1}^{r} x_i)} \qquad ...(2.15)$$

$$i = 1, 2, \ldots\ldots (r+1)$$

$$\hat{p}^2_{r+1} + \beta\, p_{r+1} = \frac{(C-1)\,(C-2)}{\left(C-1+\sum_{i=1}^{r} x_i\right)\left(C-2+\sum_{i=1}^{r} x_i\right)} + \beta\, \frac{(C-1)}{\left(C-1+\sum_{i=1}^{r} x_i\right)} \qquad \ldots(2.16)$$

$$\hat{p}_i\, \hat{p}_j = \frac{x_i\, x_j}{\left(C-1+\sum_{i=1}^{r} x_i\right)\left(C-2+\sum_{i=1}^{r} x_i\right)} \qquad i \neq j = 1, 2, \ldots\ldots r \qquad \ldots(2.17)$$

$$\hat{p}_i\, \hat{p}_{r+1} = \frac{x_i\,(C-1)}{\left(C-1+\sum_{i=1}^{r} x_i\right)\left(C-2+\sum_{i=1}^{r} x_i\right)} \qquad i = 1, 2, \ldots r$$

The unbiased estimators of $p_i^2 + \beta\, p_i$ for i = 1, 2 under BSP and IBSP are the special cases of (2.13) and (2.15) for r= 1.

Variances of $p_2^2 + \beta\, p_2$ under BSP and IBSP have been derived in Dutta [1]. We quote those here for the reference.

(v) $$V_B\,(\hat{p}_2^2 + \beta\,\hat{p}_2) = V_B\,(\hat{p}_2^2) + 2\beta\ \mathrm{cov}_B\,(\hat{p}_2^2, \hat{p}_2) + \beta^2\ V_B\,(\hat{p}_2)$$

$$= (p_2/n)\ [(\beta+2)^2\, p_1 - 2\,\{4 - 2\beta + (1/(n+1)\}\, p_1^2 + 2\ \{2 - (1/(n-1)\}\, p_1^3] \qquad \ldots(2.18)$$

where V_B stands for the variance of estimators under BSP.

(vi) Similarly,

$$V_I\,(\hat{p}_2 + \beta\,\hat{p}_2) = V_I\,(\hat{p}_2^2) + 2\beta\ \mathrm{cov}_I\,(\hat{p}_2^2, \hat{p}_2) + \beta^2\, V_I\,(\hat{p}_2)$$

$$= (p_2^2/C)\ [(\beta+2)^2\, p_1 + 2\ \frac{\beta^2 - \beta\,(C-2) - (4C-5)}{(C+1)}\ p_1^2 + 2\ \frac{3\beta^2 - 6\,\beta\,(C-2) + 2(C-2)\,(C-4)}{(C+1)\,(C+2)}\ p_1^3$$

$$+ \sum_{y=4}^{\infty} \frac{C!\,(y-1)!}{(C+y-1)!} [\{\beta^2 - 2\beta\,(C-2)\,(C-3)\}\; y - (C-1)\,(C-2)]\; p_1^y$$

...(2.19)

3. Sequential estimation of gene, genotype and phenotype proportions

Let $Y = (e_1, e_2, \ldots e_{r+1})$ be a vector of (r+1) elements where $e_i = 1$ if a particular i-th phenotype or genotype or gene is observed of an individual or of a plant or of an animal and $e_j = 0$ for $j \neq i = 1,2,\ldots(r+1)$. Hence $\sum_{i=1}^{N} Y_i = (x_1, x_2, \ldots x_{r+1})$ is a bi-variate random vector if r= 1 and multivariate random vector if $r \geq 2$. Accordingly $(x_1, x_2, \ldots x_{r+1})$ is a boundary point of two - dimensional sampling plan for a desired stopping rule when r = 1 and a boundary point of a multidimensional sampling plan for a relevant stopping rule [Ref. Dutta and Goswami [2]] Thus for every observation, y is a vector (or two vectors) of 'o' or '1' elements such that only the i-th element is 1 if the corresponding phenotype or genotype (or gene) is observed (see examples in case I). We have discussed below the problem of sequential estimation of parameters under two situations, viz. (i) when the forces which influence the change of gene frequency of a population are present and (ii) when the forces which influence the change of gene frequency of a population are absent.

CASE I — Estimation of genetic parameters when the forces which influence the change of gene frequency of a population are present.

The Hardy-Weinberg law does not hold good when the forces which influence the change of gene frequency of a population are present. In this case, the genotype proportions and phenotype proportions can be estimated (i) if the counting of gene is possible and (ii) if the counting of genotypes or phenotypes are possible.

(a) When counting of gene is possible, the total number of gene (allele) counts will be 2n if we observe n individuals as they occur in pairs. Here, the sequential variable $Y = (e_1, e_2, \ldots, e_{r+1})$ will take two values when a specific genotype is observed.

For example, consider the MN blood group system. If MM blood group is observed then the corresponding observable random vectors are (1, 0) and (1, 0), if MN blood group is observed then (1, 0) and (0, 1) are the vectors of observation and if NN blood group is observed then (0, 1) and (0, 1) are the observable random vectors. The estimation of gene proportions can be made

adopting the Binomial Sampling Plan if there is no rare gene and Inverse Binomial Sampling Plan if a specific gene is rarest one.

Secondly consider a case of three alleles, say A, B, O blood group system. In this case r = 2 and Y = (1, 0, 0) if A is observed, Y = (0, 1, 0) if B is observed and Y = (0, 0, 1) if O is observed. Hence the genotype observation AA will be converted to two observed vectors as (1,0,0) and (1,0,0), for the genotype AO as (1, 0, 0) and (0, 0, 1), for the genotype BB as (0, 1, 0) and (0, 1, 0), for the genotype BO as (0, 1, 0) and (0, 0, 1), for the genotype OO as (0, 0, 1) and (0, 0, 1) and for the genotype AB as (1, 0, 0) and (0, 1, 0). The relevant proportions of genes A,B and O are respectively p, q and r, can be estimated using Trinomial Sampling Plan if there is no rare gene and Inverse Trinomial Sampling Plan if there is a rare gene. The variance of the estimators are quadratic functions of proportions. These can be estimated using the relevant sampling plan.

Again, the genotype and phenotype proportions can be estimated adopting the same principle. For A, B, O blood group system we will have genotype as AA, AO, BB, BO, OO and AB. If AA is observed then the observed vector for it is (1,0,0,0,0,0), if AO is observed, then the observed vector for it is (0,1,0,0,0,0), if BB is observed then the observed vector is (0,0,1,0,0,0) and so on.

The phenotic observations will be 'A' group for AA, AO, 'B' group for BB, BO, 'O' group for OO and 'AB' group for AB. In this case the observable vector will have four components with '1' in one place and zero in rest. The Multinomial and the Inverse Multinomial Sampling Plans can be adopted for the estimation of proportions and the variances of estimators assuming r = 3.

In general, if we have (r+1) alleles then each observations can be represented as the vector Y of (r+1) elements. Suppose, the ith allele A_i is observed then $e_i = 1$ and $e_j = 0$ for $j \neq i = 1, 2,.. (r+1)$. If A_iA_j is the observed genotype, the corresponding observational vectors are (0, 0, ..., e_i = 1,..., 0, 0) and (0, 0, ..., e_j=1, ..., 0, 0). The estimation of corresponding gene proportions can be made using MSP if there is no rare gene or genotype proportions and the IMSP if there is rare gene or the genotype proportions. We have discussed in the appendix a practical example for the estimation of "Rh-" blood group.

Case-II : Estimation of genetic parameters when the forces which influence the change of gene frequency of a population are absent

Hardy-Weinberg law holds when the forces which influence the change of gene frequency of a population are absent. If we have two alleles at a locus,

then the genotype distribution of the allelic combination is $p^2.2p(1-p)$ and $(1-p)^2$ but the phenotype distribution is $p^2+2p(1-p)$ and $(1-p)^2$. In the above cases, we will assume $p_1 = p^2$, $p_2 = 2p(1-p)$, $p_3 = (1-p)^2$ to estimate genotype proportions and $p_1 = p^2 + 2p(1-p)$, $p_2 = (1-p)^2$ to estimate phenotype proportions. We use three dimensional sampling plan to estimate genotype proportions and two dimensional sampling plans to estimate phenotype proportions. If a proportion is very small, inverse binomial sampling plan should be used to estimate all proportions. We will use the estimator $(1-p)$ as $\sqrt{\hat{p}_3}$ and p as $1-\sqrt{\hat{p}_3}$ in first case. The estimator of (1+p) as $\sqrt{\hat{p}_2}$ and p as $1-\sqrt{\hat{p}_2}$ in second case. If p_1 or p_3 is small, then inverse trinomial sampling plan is applicable to estimate the rarest genotype in first case. If p_2 is very small in second case, then the Inverse Binomial Sampling Plan is used to estimate the rarest phenotype. The variances of the estimators are quadratic functions of p_1, p_2 for r = 1 and p_1, p_2, p_3 for r = 2. The conversion of genetic observations to the Bernoullian form is same as discussed in Case I. If the counting of gene is possible, then the estimation of gene, genotype proportions can be made as discussed in Case I.

Appendix

We know that in blood group sample the "Rh-" is the rarest one. In this case if "Rh+" is observed then the observational vector will be (1,0) and if "Rh-" is observed then the observational vector will be (0,1). Suppose, our stopping rule is that "continue taking observations till we get C = 7 observations of 'Rh-". Let us assume that we stop after having 108 observations as per rule. Then $\sum_{i=1}^{n} Y_i = (X_1, X_2)$ where $X_1 = 101$ and $X_2 = C = 7$. The estimators of proportions are $(x_1/C+x_1-1,\ C-1/C+x_1-1)$. The variances of estimators can be computed as stated in section-1.

ACKNOWLEDGEMENT

The authors express their deep gratitude to Prof. Prem Narain for his valuable suggestions in drafting the manuscript.

REFERENCES

[1] Dutta, K., 1995. Estimation of quadratic functions of the Bernoulli parameter in inverse sampling method. Accepted for publication in the *J. Statist. Plan. Inference.*

[2] Dutta, K. and Goswami, V., 1994. A comparative study for multidimensional sampling plans. *Jour. Ind. Soc. Ag. Stat.*, **46(2)**, 254-266.

[3] Gardner, Eldon. J., 1972. Principle of Genetics. Third edition, Wiley Pvt. Ltd.

[4] Narain, P., 1990. Statistical Genetics. John Wiley and Sons, New York.