Journ. Japan Statist. Soc.
9, 1, 1979, 29~35

ESTIMATION OF THE DIFFERENCE BETWEEN TWO PROPORTIONS IN A FINITE MULTINOMIAL POPULATION

Yoshiaki Funatsu

In a finite multinomial population, the method of interval estimation for specified single proportion is already well known. It can be applied to the estimation for sum of specified two or more proportions. However, the joint estimation for specified two proportions or estimation for the difference between the two are more complicated than that for single proportion.

This paper provides the confidence regions for specified two proportions and the difference between the two with a degree of confidence coefficient. The confidence coefficient is given below by the minimum value instead of proper one that can be found by further study.

1. Introduction

Let us consider a multinomial population Π consisting of N units which is divided into three mutually exclusive classes Π_α, Π_β and Π_γ consisting of N_α, N_β and N_γ units respectively. Let p, q and r denote the proportions of units included in Π_α, Π_β and Π_γ in Π respectively, i.e.,

$$p=\frac{N_\alpha}{N}, \quad q=\frac{N_\beta}{N}, \quad r=\frac{N_\gamma}{N},$$

$$0\leqq p\leqq 1, \quad 0\leqq q\leqq 1, \quad 0\leqq r\leqq 1, \quad p+q+r=1$$

Now, if a simple random sample of size n is drawn without replacement out of this population, we can divide the n units into three mutually exclusive classes Π_α, Π_β and Π_γ consisting of n_α, n_β and n_γ respectively. The probability that a set of three values n_α, n_β and n_γ occurs is given

$$\Pr(n_\alpha, n_\beta, n_\gamma)=\frac{\binom{N_\alpha}{n_\alpha}\binom{N_\beta}{n_\beta}\binom{N_\gamma}{n_\gamma}}{\binom{N}{n}}$$

where

$$0\leqq n_\alpha\leqq\min(n, N_\alpha), \quad 0\leqq n_\beta\leqq\min(n, N_\beta), \quad 0\leqq n_\gamma\leqq\min(n, N_\gamma)$$

$$n_\alpha+n_\beta+n_\gamma=n, \qquad N_\alpha+N_\beta+N_\gamma=N.$$

Let $\hat{p}$, $\hat{q}$ and $\hat{r}$ denote the sample proportions of units included in Π_α, Π_β and Π_γ in Π respectively, i.e.,

$$\hat{p}=\frac{n_\alpha}{n}, \quad \hat{q}=\frac{n_\beta}{n}, \quad \hat{r}=\frac{n_\gamma}{n},$$

Received Aug. 16, 1978, Revised Jan. 19, 1979.

$$0\leqq\hat{p}\leqq 1\,,\quad 0\leqq\hat{q}\leqq 1\,,\quad 0\leqq\hat{r}\leqq 1\,,\quad \hat{p}+\hat{q}+\hat{r}=1\,.$$

Further, let $\hat{p}'$, $\hat{q}'$ and $\hat{r}'$ denote the ratios of n_α, n_β and n_γ to N respectively, i.e.,

$$\hat{p}'=\frac{n_\alpha}{N}\,,\quad \hat{q}'=\frac{n_\beta}{N}\,,\quad \hat{r}'=\frac{n_\gamma}{N}\,,\quad \hat{p}'+\hat{q}'+\hat{r}'=\frac{n}{N}\,.$$

It is obvious that these three ratios $\hat{p}'$, $\hat{q}'$ and $\hat{r}'$ become fixed values after a sample of size n was drawn, and always smaller than p, q and r respectively. Consequently we have following three inequalities that restrict the values of p and q.

(1) $$p\geqq\hat{p}'\,,\quad q\geqq\hat{q}'\,,\quad p+q\leqq 1-\hat{r}'$$

2. Confidence region for p and q

Let us denote by

$$d=p-q$$
$$\hat{d}=\hat{p}-\hat{q}$$

It is known that the expectation and variance of $\hat{d}$ are expressed as follows.

(2) $$\begin{cases} E(\hat{d})=d \\ V(\hat{d})=\dfrac{N-n}{N-1}\cdot\dfrac{p+q-d^2}{n}\,. \end{cases}$$

Combining these formulas by Tchebychef's inequality for $\hat{d}$, we obtain

$$\Pr\{|\hat{d}-d|<\lambda\sqrt{V(\hat{d})}\}\geqq 1-\frac{1}{\lambda^2}\,,\qquad \lambda>0\,.$$

More generally, this can be written

(3) $$\Pr\{|\hat{d}-d|\leqq\lambda\sqrt{V(\hat{d})}\}=1-w\geqq 1-\frac{1}{\lambda^2}\,,\quad \lambda>0\,,\quad 0\leqq w\leqq\frac{1}{\lambda^2}$$

since

$$\Pr\{|\hat{d}-d|\leqq\lambda\sqrt{V(\hat{d})}\}\geqq\Pr\{|\hat{d}-d|<\lambda\sqrt{V(\hat{d})}\}$$

Using (2) and (1), (3) is expressed in detail as follows:

(4) $$\begin{cases} \Pr\{|\hat{d}-d|\leqq\sqrt{t(p+q-d^2)}\}=1-w\geqq 1-\dfrac{1}{\lambda^2} \\ p\geqq\hat{p}'\,,\quad q\geqq\hat{q}'\,,\quad p+q\leqq 1-\hat{r}' \end{cases}$$

where

$$t=\frac{N-n}{N-1}\cdot\frac{\lambda^2}{n}\,,\quad \lambda>0\,,\quad 0\leqq w\leqq\frac{1}{\lambda^2}\,.$$

Value of w will be approximately 0.05 for $\lambda=2$ and 0.003 for $\lambda=3$ respectively, since $\hat{d}$ is asymptotically normally distributed, if N and n are large, n/N is small, p and q are not very close to zero or unity.

Now we shall try to find a region that includes the parameter $\theta=\left(\dfrac{N_\alpha}{N},\dfrac{N_\beta}{N}\right)$

with an appraisal of probability.

Denoting by

Θ: Space of parameter θ

I_θ: Space of interval $[d-\sqrt{t(p+q-d^2)}\ ,\ d+\sqrt{t(p+q-d^2)}]$, where $p\geqq\hat{p}'$, $q\geqq\hat{q}'$, $p+q\leqq 1-\hat{r}'$

ξ: Space of $\hat{d}$

(4) can also be expressed as

$$\Pr(\hat{d}\in I_\theta)\geqq 1-\frac{1}{\lambda^2}$$

where $\hat{d}\in I_\theta$ means that $\hat{d}$ is covered by I_θ. $\hat{d}$ is not an element of I_θ.

Let us consider a set S defined by

$$S=\{(\theta,\xi):\ \theta\in\Theta,\ \xi\in I_\theta\}$$

in the three dimensional space, and denote by A the plane of intersection of the set S and a plane $\xi=\hat{d}$, then the following three events are equivalent.

$$\hat{d}\in I_\theta,\quad (\theta,\hat{d})\in S,\quad \theta\in \mathrm{A}\ .$$

It follows that

$$\Pr(\theta\in \mathrm{A})=\Pr(\hat{d}\in I_\theta)\geqq 1-\frac{1}{\lambda^2}\ .$$

Hence we see A is a confidence region for θ with a minimum value of confidence coefficient $1-\frac{1}{\lambda^2}$ (exactly saying, the region consists of all "lattice" points $\left(\frac{x_\alpha}{N},\frac{x_\beta}{N}\right)$ in A where x_α and x_β are non-negative integers under N).

The locus of A can explicitly be shown by transforming (4). First, solving q in the inequality in the braces of (4), it is easily transformed to

$$q_1\leqq q\leqq q_2$$

where

(5)
$$\begin{cases} q_1=\dfrac{2(1+t)p-(2\hat{d}-t)-\sqrt{G}}{2(1+t)} \\ q_2=\dfrac{2(1+t)p-(2\hat{d}-t)+\sqrt{G}}{2(1+t)} \end{cases}$$

$$G=(2\hat{d}-t)^2-4(1+t)(\hat{d}^2-2tp)\ .$$

The loci of q_1 and q_2 form a parabola whose axis makes angles of 45° with the two coordinate axes (See Fig. 1).

Let A′ denote the region enclosed by the parabola mentioned above, and A″ the region enclosed by the three straignt lines $p=\hat{p}'$, $q=\hat{q}'$, $p+q=1-\hat{r}'$ that form a right triangle with two equal sides. A, the confidence region for θ, is the intersection of A′ and A″, i.e.,

$$\mathrm{A}=\mathrm{A}'\cap\mathrm{A}''\qquad \text{(See Fig. 1).}$$

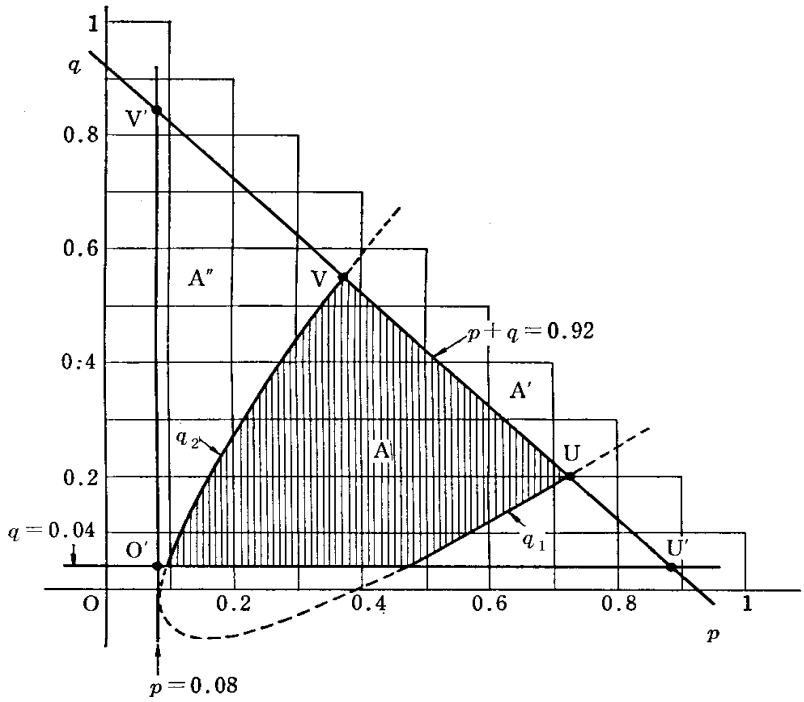

Fig. 1 Locus of A

A : Region defined by
(i) $|\hat{d}-d| \leqq \sqrt{t(p+q-d^2)}$
(ii) $p \geqq \hat{p}'$, $q \geqq \hat{q}'$, $p+q \leqq 1-\hat{r}'$
where $\hat{d}$, t, $\hat{p}'$, $\hat{q}'$ and $\hat{r}'$ are given.

A′ : Region defined by (i), where $\hat{d}$ and t are given (Region enclosed by parabola).

A″: Region defined by (ii), where $\hat{p}'$, $\hat{q}'$ and $\hat{r}'$ are given (Region enclosed by right triangle O′U′V′).

$A=A' \cap A''$

(Numerical example)

$N=100$, $n=20$, $\lambda=2$, $t=0.\dot{1}\dot{6}$ $\quad$ $\hat{p}=0.4$, $\hat{q}=0.2$, $\hat{r}=0.4$, $\hat{d}=0.2$
$n_\alpha=8$, $n_\beta=4$, $n_\gamma=8$ $\quad$ $\hat{p}'=0.08$, $\hat{q}'=0.04$, $\hat{r}'=0.08$

3. Confidence region for d

The difference $d=p-q$ takes various values in A. Clearly, in A, the minimum value and the maximum value of d occur on the straight line $p+q=1-\hat{r}'$. For an interval estimation for d, we denote by

$d(\mathrm{V})$: Value of d at the point V in Fig. 2,

d_1 : The minimum value of d in A,

d_2 : The maximum value of d in A.

It follows that

$$d_1 \leqq d \leqq d_2 \tag{6}$$

in A.

From the properties of the two regions A′ and A″, we can see

$$\begin{cases} d_1=\max\{d(\mathrm{V}),\ d(\mathrm{V}')\} \\ d_2=\min\{d(\mathrm{U}),\ d(\mathrm{U}')\} \end{cases} \tag{7}$$

A= ||||||||||| B= ≡≡≡

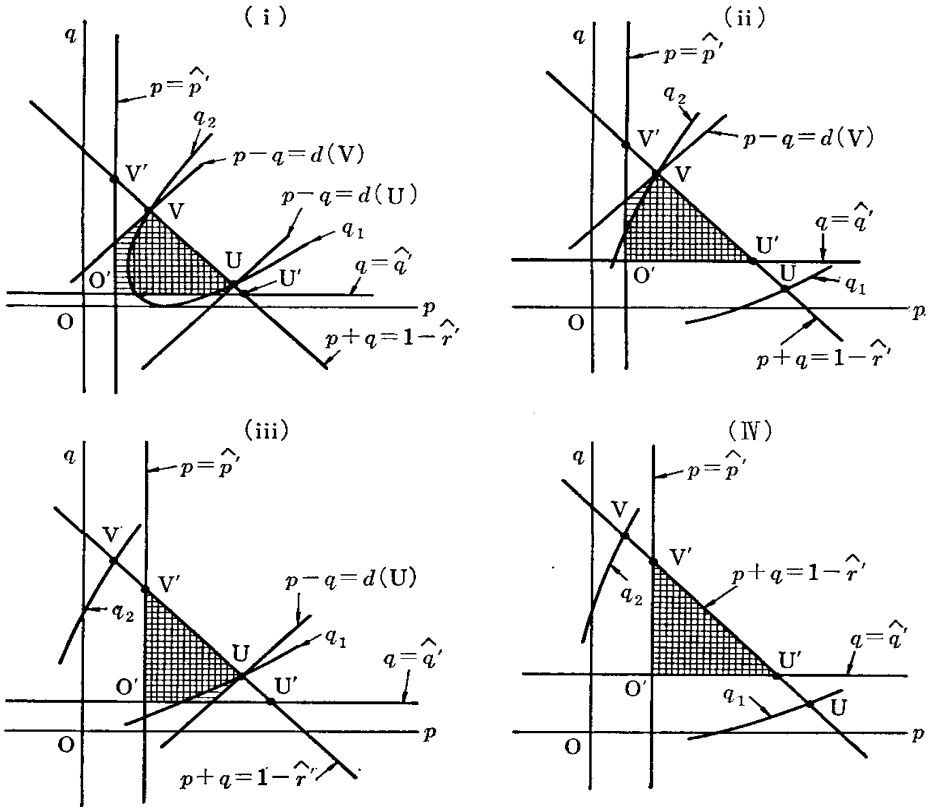

Fig. 2 Various cases of the regions A and B

U : Point of intersection of a straight line $p+q=1-\hat{r}'$ and a curve q_1.

U′: Point of intersection of two straight lines $p+q=1-\hat{r}'$ and $q=\hat{q}'$.

V : Point of intersection of a straight line $p+q=1-\hat{r}'$ and a curve q_2.

V′: Point of intersection of two straight lines $p+q=1-\hat{r}'$ and $p=\hat{p}'$.

O′: Point of intersection of two straight lines $p=\hat{p}'$ and $q=\hat{q}'$.

where $d(\mathrm{V})$ and $d(\mathrm{U})$ are obtained as follows by solving d at the formulas (5) with $p+q=1-\hat{r}'$, i.e.,

$$(8)\qquad \begin{cases} d(\mathrm{V})=\dfrac{\hat{d}-\sqrt{t\{1-\hat{r}'+t(1-\hat{r}')-\hat{d}^2\}}}{1+t} \\ d(\mathrm{U})=\dfrac{\hat{d}+\sqrt{t\{1-\hat{r}'+t(1-\hat{r}')-\hat{d}^2\}}}{1+t} \end{cases}.$$

And more easily we obtain

$$(9)\qquad \begin{cases} d(\mathrm{V}')=-(1-2\hat{p}'-\hat{r}') \\ d(\mathrm{U}')=1-2\hat{q}'-\hat{r}' \end{cases}.$$

(6) will be correct in some cases and wrong in some cases, since d_1 and d_2 are

variable. In order to obtain the probability that (6) is correct, we shall consider a region B whose area is defined by the following inequalities (See Fig. 2).

$$(10)\qquad \begin{cases} p+q\leqq 1-\hat{r}' \\ p\geqq \hat{p}' \\ q\geqq \hat{q}' \\ \text{If } d_1=d(\mathrm{V}), \text{ then } p-q\geqq d(\mathrm{V}) \text{ is added.} \\ \text{If } d_2=d(\mathrm{U}), \text{ then } p-q\leqq d(\mathrm{U}) \text{ is added.} \end{cases}$$

Clearly, B completely includes A, that is, $\mathrm{A}\subset\mathrm{B}$. Consequently, if A includes the point $\theta=\left(\frac{N_\alpha}{N}, \frac{N_\beta}{N}\right)$, B certainly includes it. But converses are not always true. Hence we obtain

$$(11)\qquad \Pr(\theta\in \mathrm{B})\geqq \Pr(\theta\in \mathrm{A})=1-w\geqq 1-\frac{1}{\lambda^2}\,.$$

Further, we can see that if (6) is correct, the event in the parentheses of left side of (11) is also correct and the converses are true, too. This means the two events $d_1\leqq d\leqq d_2$ and $\theta\in \mathrm{B}$ are equivalent each other. Hence we finally arrive the following conclusion.

$$(12)\qquad \Pr(d_1\leqq d\leqq d_2)\geqq 1-w\geqq 1-\frac{1}{\lambda^2}$$

4. A numerical example

If $N=100$, $n=20$, $\lambda=2$, $n_\alpha=8$, $n_\beta=4$, $n_\gamma=8$, then

$$\hat{p}=\frac{8}{20}=0.4\,,\quad \hat{q}=\frac{4}{20}=0.2\,,\quad \hat{d}=0.4-0.2=0.2\,,\quad \hat{p}'=\frac{8}{100}=0.08\,,$$

$$\hat{q}'=\frac{4}{100}=0.04\,,\quad \hat{r}'=\frac{8}{100}=0.08\,,\quad t=\frac{100-20}{100-1}\cdot\frac{2^2}{20}=0.\dot{1}\dot{6}\,.$$

The locus of A is sketched in Fig. 1.

From (8) and (9),

$$d(\mathrm{V})=\frac{0.2-\sqrt{0.\dot{1}\dot{6}\{1-0.08+0.\dot{1}\dot{6}(1-0.08)-0.2^2\}}}{1+0.\dot{1}\dot{6}}=-0.1788\,,$$

$$d(\mathrm{U})=\frac{0.2+\sqrt{0.\dot{1}\dot{6}\{1-0.08+0.\dot{1}\dot{6}(1-0.08)-0.2^2\}}}{1+0.\dot{1}\dot{6}}=0.5232\,,$$

$$d(\mathrm{V}')=-(1-2\times 0.08-0.08)=-0.76\,,$$

$$d(\mathrm{U}')=1-2\times 0.04-0.08=0.84\,.$$

From (7),

$$d_1=\max(-0.1788,\ -0.76)=-0.1788\,,$$

$$d_2=\min(0.5232,\ 0.84)=0.5232\,.$$

From (12),

$$\Pr(-0.1788 \leqq d \leqq 0.5232) \geqq 0.75$$

Since $d=\frac{N_\alpha}{N}-\frac{N_\beta}{N}$ where N_α and N_β are integers under N, it is sufficient in this case to express as

$$\Pr(-0.17 \leqq d \leqq 0.52) \geqq 0.75 .$$

Acknowledgement

The author is deeply grateful to the referees for their comments without which this work would not be successful.