Determination of Sample Size in the Mixture of Normal Distributions

メタデータ	言語: eng 出版者: 公開日: 2010-04-06 キーワード (Ja): キーワード (En): 作成者: Koike, Toshitaka, Mori, Ken'ichi, Kase, Shigeo メールアドレス: 所属:
URL	https://doi.org/10.24729/00008602

Determination of Sample Size in the Mixture of Normal Distributions

Toshitaka KOIKE*, Ken'ichi MORI ** and Shigeo KASE**

(Received Oct. 29, 1982)

Two-sample scheme by Stein has been utilized to determine the practical sample size in many opportunities. Since he assumes that the population is normal, his technique can not be applied directly to non-normal population. The present paper deals with the determination of sample size applicable in the case where the population is a mixture of normal distributions. The determination method using range is selected for modification because of simplicity of calculation. The estimation procedure of variance of mixed normal distribution is derived through estimation of expectation of the range. The robustness of Stein's method in this population is also discussed.

1. Introduction

Since Stein[4)] devised a two-stage sample size determination scheme, many authors have investigated this procedure and given some derivatives from this scheme. However, the Stein's method and other's assume the normal population from which the sample is drawn. If the sampled population is not follow the normal distribution, change of sample size necessary to estimate the population parameters becomes a serious matter of great concern.

The method which uses a sample range after Stein's has been utilized from its simplicity of calculation. This paper discusses a sample size determination method with a range for the population which is a mixture of two normal distributions with different means. Specifically, the expected value of sample range in mixed normal distribution is first treated, and thereby an estimating method of variance in the mixed normal distribution is proposed. One of the sample size determination method in the mixed normal distribution is also discussed.

2. Stein's Sample Size Determination Procedure

Stein's two-sample procedure for estimating the mean of normal population $N(\mu, \sigma^2)$ with a confidence interval of preassigned length is summarized as follows :

Draw the first sample of tentative size N_1, and calculate

$$\bar{x}_1 = \sum_{i=1}^{N_1} x_i/N_1 ,$$

$$\hat{\sigma}^2 = \sum_{i=1}^{N_1} (x_i - \bar{x}_1)^2 / (N_1 - 1) .$$

With the above values, determine the necessary sample size N by

$$N = (t_{N_1-1} (\alpha)\, \hat{\sigma} / l)^2 , \qquad (1)$$

* Ryukoku University, Faculty of Business Administration.

** Department of Industrial Engineering, College of Engineering.

where $t_{N_1-1}(\alpha)$ is the two sided α point of Student's t-distribution with N_1-1 degrees of freedom and $2l$ stands for the length of confidence interval given in advance.

Some other investigators derive the simplified method using sample range r by dint of this procedure. The unbiassed estimate of variance is given by $\hat{\sigma}^2 = (r/d_2)^2$, where sample range $r = x_{(n)} - x_{(1)}$. Substitution of this into the Stein's method yields the sample size such as

$$N = (t_{N_1-1}(\alpha)\, r/(d_2 l))^2, \qquad (2)$$

where d_2 is the expected value of the distribution of a range from standard normal distribution, d_2's are already tabulated for use in control charts, and $x_{(i)}$ stands for the order statistic in the sample of size n.

Replacement of $\bar{x}_1/40$ for $l/t_{N_1-1}(\alpha)$ reduces[3] Eq. (2) to

$$N = (40\, r/(d_2 \bar{x}_1))^2. \qquad (3)$$

This treatment means that confidence coefficient is about equal to 0.95 and the length of confidence interval is about 5 percent of $\bar{x}_1$.

3. Distribution of Order Statistics from Mixed Population

3.1 Distribution of Symmetric Statistic

Behboodian[1] shows the distribution of symmetric statistic $T = g(X_1, X_2, \ldots, X_n)$ from the mixed population which is composed of two probability density functions $f_1(x)$ and $f_2(x)$ with mixing proportions p ($0 < p < 1$) and q such that

$$f(x) = p f_1(x) + q f_2(x), \qquad q = 1 - p.$$

He points out that the density $f_T(t)$ of symmetric statistic T is given by

$$f_T(t) = \sum_{k=0}^{n} {}_nC_k\, p^k q^{n-k} f_{T_k}(t), \qquad (4)$$

that is, the density of T is a binomial mixture of the densities $f_{T_k}(t)$ of the T_k's. $T_k = g(X_{k1}, X_{k2}, \ldots, X_{kn})$ and T_k is a statistic for which the X_{ki}'s are independent with density $f_1(x)$ if $i \leqq k$ and density $f_2(x)$ if $i > k$. Moreover, symmetric statistic of a random sample stands for the statistic $T = g(X_1, X_2, \ldots, X_n)$ which is invariant by any permutation on X_i's.

3.2 Distributions of Order Statistics

Since the order statistic is one of symmetric statistics, its distribution from mixed population can be obtained as described in 3.1. From Eq. (4) the first order statistic $X_{(1)}$ in ascending order for variables $X_1, X_2, \ldots, X_n$ has probability density

$$f_{x_{(1)}}(x) = \sum_{k=0}^{n} {}_nC_k\, p^k\, q^{n-k} f_{x_{(k1)}}(x), \qquad (5)$$

where $X_{(k1)}$ is the first order statistic in the sample $X_{k1}, X_{k2}, \ldots, X_{kn}$. The density of $X_{(k1)}$, $f_{X_{(k1)}}(x)$, is given as

$$f_{X_{(k1)}}(x) = k f_1(x) \left[1-F_1(x)\right]^{k-1} \left[1-F_2(x)\right]^{n-k}$$

$$+ (n-k) f_2(x) \left[1-F_1(x)\right]^{k} \left[1-F_2(x)\right]^{n-k-1}, \tag{6}$$

where $F_i(x)$, $i = 1, 2$, is the distribution function. The distributions of the other order statistics can be similarly obtained.

4. Sample Size Determination for Mixed Distribution

4.1 Expected Value of Range from Mixed Normal Distribution

Consider the range R of sample from mixed normal distribution

$$f(x) = (p/\sigma)\,\phi[(x-\mu_1)/\sigma] + (q/\sigma)\,\phi[(x-\mu_2)/\sigma],$$

which results from mixing of two normal distributions having means μ_1, μ_2, respectively, and equal variance σ^2, where $\phi(t)$ is a density of standard normal distribution. Since $R = X_{(n)} - X_{(1)}$, the expectation of R becomes

$$E(R) = E(X_{(n)} - X_{(1)}) = E(X_{(n)}) - E(X_{(1)}).$$

The distribution of $X_{(1)}$ from mixed normal distribution is given by Eqs. (5) and (6), and the one of $X_{(n)}$ is calculated in the same manner described in 3.2 as

$$f_{X_{(n)}}(x) = \sum_{k=0}^{n} {}_nC_k\, p^k q^{n-k} f_{X_{(kn)}}(x), \tag{7}$$

$$f_{X_{(kn)}}(x) = (k/\sigma)\,\phi[(x-\mu_1)/\sigma] \left\{\Phi[(x-\mu_1)/\sigma]\right\}^{k-1}$$

$$\left\{\Phi[(x-\mu_2)/\sigma]\right\}^{n-k} + ((n-k)/\sigma)\,\phi[(x-\mu_2)/\sigma] \left\{\Phi[(x-\mu_1)/\sigma]\right\}^{k}$$

$$\left\{\Phi[(x-\mu_2)/\sigma]\right\}^{n-k-1}, \tag{8}$$

where $\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt$. Therefore, $E(R)$ may be rewritten as

$$E(R) = \sum_{k=0}^{n} {}_nC_k\, p^k q^{n-k} E(R_k), \tag{9}$$

where $E(R_k)$ stands for

$$E(R_k) = E(X_{(kn)}) - E(X_{(k1)}). \tag{10}$$

From Eqs. (6), (8), and (9), it is easily obtained by the same derivation as in the normal case that

$$E(R_k) = \int_{-\infty}^{\infty} x \left\{f_{X_{(kn)}}(x) - f_{X_{(k1)}}(x)\right\} dx$$

$$= \sigma \int_{-\infty}^{\infty} \left[1 - \{\Phi(x)\}^k \{\Phi(x+\Delta)\}^{n-k} - \{1-\Phi(x)\}^k \right.$$

$$\left. \{1-\Phi(x+\Delta)\}^{n-k}\right] dx$$

$$\equiv \sigma D(k), \tag{11}$$

where $\Delta = (\mu_1 - \mu_2) / \sigma$. Consequently Eq. (9) is rewritten as

$$E(R) = \sum_{k=0}^{n} {}_nC_k \, p^k \, q^{n-k} \, \sigma D(k) \equiv \sigma D \ . \tag{12}$$

The coefficient D is defined by

$$D = \sum_{k=0}^{n} {}_nC_k \, p^k \, q^{n-k} \, D(k) \ , \tag{13}$$

which is a function of parameters n, p and Δ .

If there exists no mixture of distribution, we may use $\Delta = 0$ or $p = 0$ (or $p = 1$) in Eqs. (11) and (12). In this case, it is readily known that Eq. (13) gives the same coefficient D as in the normal case d_2. Table 1 shows the coefficients D's for some values of n, p and Δ. This table gives the D's for p ranging up to 0.5. The D's for p more than 0.5 are symmetrical with respect to $p = 0.5$, so that any value of D can be easily obtained from this relation.

Table 1 Table of coefficients D

	Δ \ p	0.1	0.2	0.3	0.4	0.5
$n = 10$	0.5	3.112	3.139	3.157	3.168	3.172
	1.0	3.215	3.315	3.382	3.421	3.434
	1.5	3.383	3.588	3.719	3.791	3.815
	2.0	3.606	3.933	4.128	4.231	4.263
	2.5	3.872	4.325	4.578	4.705	4.744
	3.0	4.166	4.746	5.048	5.193	5.236
	3.5	4.478	5.181	5.528	5.687	5.733
	4.0	4.797	5.623	6.012	6.183	6.232
	4.5	5.121	6.068	6.497	6.680	6.731
	5.0	5.446	6.514	6.983	7.177	7.230
$n = 20$	0.5	3.777	3.809	3.831	3.844	3.849
	1.0	3.907	4.023	4.098	4.140	4.154
	1.5	4.124	4.352	4.485	4.555	4.577
	2.0	4.422	4.762	4.939	5.026	5.052
	2.5	4.799	5.217	5.422	5.518	5.546
	3.0	5.178	5.696	5.917	6.016	6.045
	3.5	5.598	6.184	6.415	6.515	6.544
	4.0	6.029	6.676	6.914	7.015	7.044
	4.5	6.466	7.170	7.413	7.515	7.544
	5.0	6.904	7.664	7.913	8.015	8.044
$n = 30$	0.5	4.132	4.167	4.191	4.205	4.209
	1.0	4.277	4.401	4.478	4.521	4.535
	1.5	4.523	4.757	4.887	4.953	4.974
	2.0	4.861	5.193	5.355	5.433	5.457
	2.5	5.264	5.668	5.846	5.929	5.953
	3.0	5.707	6.159	6.344	6.428	6.453
	3.5	6.170	6.656	6.843	6.928	6.953
	4.0	6.643	7.155	7.343	7.427	7.452
	4.5	7.119	7.654	7.843	7.927	7.952
	5.0	7.598	8.153	8.343	8.428	8.452

4.2 Determination of Sample Size

If a population distribution is not normal but mixed normal, the Stein's sample size determination method, Eqs. (2) and (3), can not give the correct value in the original form. In what follows, the device in such a circumstance shall be discussed.

If p and Δ of mixed normal are known, the estimate $\hat{\sigma}^2$ of σ^2 is given by

$$\hat{\sigma}^2 = (r/D)^2 = \{(x_{(N_1)} - x_{(1)})/D\}^2$$

from sample range r and coefficient D for $n = N_1$. However, such derived $\hat{\sigma}^2$ is not a variance of mixed distribution but an unbiassed estimate of variance of original normal distribution. The population mean θ_1' and population variance θ_2 of mixed distribution are

$$\begin{aligned} \theta_1' &= p\mu_1 + q\mu_2 , \\ \theta_2 &= \sigma^2 + p(\mu_1 - \theta_1')^2 + q(\mu_2 - \theta_1')^2 . \end{aligned} \tag{14}$$

By use of the relation of Eq. (14), the estimate of variance of mixed distribution can be derived. In order to generalize the assumption slightly, let p and μ_2 be known. The algorithm of sequential estimation of the variance is shown as follows :

1) Determine an initial value of $\hat{\sigma}$.
2) Calculate $\Delta \equiv (\hat{\mu}_1 - \mu_2)/\hat{\sigma} = (\bar{x}_1 - \mu_2)/(p\hat{\sigma})$.
3) Find the coefficient D corresponding to N_1, p and Δ from the table, and obtain the new estimate $\hat{\sigma}_1$ as $\hat{\sigma}_1 = r/D$.
4) If $|\hat{\sigma}_1 - \hat{\sigma}| < \epsilon$ (infinitesimal number), then proceed to 5), otherwise set $\hat{\sigma} = \hat{\sigma}_1$, go back to 2) and repeat.
5) From Eq. (14), the estimate of variance $\hat{\theta}_2$ becomes

$$\hat{\theta}_2 = \hat{\sigma}^2 + ((1-p)/p)(\bar{x}_1 - \mu_2)^2 . \tag{15}$$

If $\hat{\theta}_2$ is obtained by the above algorithm, the total sample size N can be calculated by

$$N = (t_{N_1-1}(\alpha)/l)^2 \, \hat{\theta}_2 , \tag{16}$$

or

$$N = (40/\bar{x}_1)^2 \, \hat{\theta}_2 , \tag{17}$$

from Eq. (2) or (3).

4.3 Effect due to Prior Informations

This section treats the changes of sample size under the condition that the values of p and μ_2 deviate fairly from their true values. Suppose that a procedure based on Eq. (16) is applied to the determination of sample size. Dependency of sample size on prior information for Eq. (17) can be discussed in the similar way as described in this section.

It is rather troublesome for one to treat the sample size analytically because of its non-linearity in p and μ_2, so we will calculate its change numerically. Let N_1 (the first sample size) be equal to 10, the length of confidence interval (2*l*) 400, and confidence coefficient (1 $-\alpha$) 0.9. Assume also that the sample range (*r*) and the sample mean ($\bar{x}_1$) are calculated from the first stage sample as 1800 and 3500, respectively. These postulated values give total sample size *N*'s for several values of *p* and μ_2 by the procedure in 4.2. Fig. 1 shows these results as contours of *N*. The value on each lattice point is *N* for the corresponding *p* and μ_2. In Fig. 1, the column of p = 1.0 and the row of μ_2 = 3500 show the values in case of no mixture. The shifts from these column and row mean that there exists the mixture. The values in case of $\mu_2 \geqq 3500$ is only shown in Fig. 1, because the another case is given by symmetry.

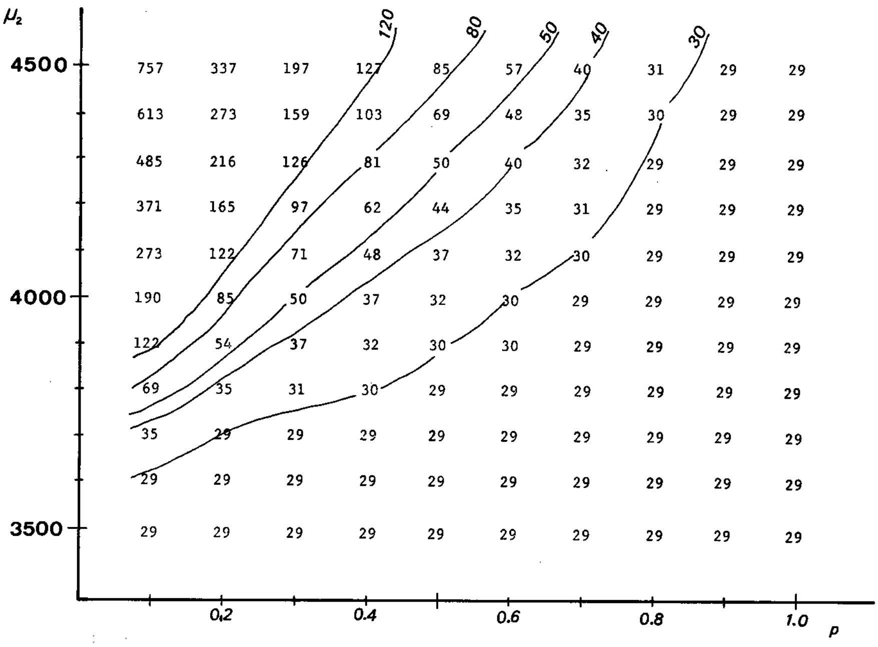

Fig. 1 The change of sample size as a function of μ_2, p

From Fig. 1, it is known that little effect by mixture is recognized when p = 0.9. Little drift of sample size is also seen in case of p = 0.8. If the mixture of distribution is limited to this extent, the shift of μ_2 from its true value scarecely has an effect on sample size. If μ_2 is nearly equal to 3500, the change of *p* brings little change of sample size. In fact, if *p* is enough large or if the means of the two distributions are not so distinct from each other, the shift of prior knowledge of *p* and μ_2 has little effect on sample size.

Contrarily if *p* is small or if there is a significant difference between the two means of distributions which are mixed, it is concluded from Fig. 1 that the sample size is quite sharply effected by the shifts of *p* and μ_2. When one is to determine the sample size in this situation, *p* and μ_2 must be estimated deliberately.

If *p* is very small and μ_2 differs from 3500 meaningfully, extraordinary sample size is calculated. For example, if *p* is 0.1 and μ_2 is 2500, then N = 757. For this instance,

μ_1 is estimated as $\hat{\mu}_1 = (3500 - 0.9 \times 2500) / 0.1 = 12500$. It is rather hard to consider that a sample range as 1800 is obtained from the mixed normal distribution with $\mu_1 = 12500$ and $\mu_2 = 2500$. The reason why an extraordinary sample size such as 757 is calculated seems to be due to this fact. If such an extreme result is encountered, one must regard the prior knowledge as wrong and reestimate these values.

4.4 Robustness of Stein's Method

Assume X_i is a random variable with mean θ_i, and the order of variance and covariance of $X_1, X_2, \ldots, X_n$ is n^{-r} ($r > 0$). Then the expectation of function $g(X_1, \ldots, X_n)$ is given[2] by

$$E(g(X_1, \ldots, X_k)) = g(\theta_1, \ldots, \theta_k) + O(n^{-r}) .$$

This relation suggests that if the population is distributed normally $N(\mu, \sigma^2)$, the expectation of Eq. (3) E_n (N) is

$$E_n(N) = [40\,\sigma/\mu]^2 .$$

On the other hand, if the population distribution is a mixed normal (mean $\theta_1{}'$ and variance θ_2), the expectation E_m (N) of Eq. (3) becomes

$$E_m(N) = [40 D\sigma / (d_2 \theta_1{}')]^2 .$$

Let $\mu = \theta_1$, $\sigma^2 = \theta_2$, then, from Eq. (14), the ratio of expectations in normal and mixed normal distribution is determined as

$$E_n(N) / E_m(N) = (d_2 / \Delta)^2 \; [1 + p(1-p)\Delta^2] . \tag{18}$$

Eq. (18) which is a function of N, p, and Δ gives the robustness of Eq. (3). Table 2 summarizes some numerical results of Eq. (18). From this table it can be seen that the ratio of sample size expectation for each N shows the same behavior against the variation of p and Δ. If $p \doteqdot 0.1$ (or $p \doteqdot 0.9$), the ratio scarcely differs from unity even for large Δ, and this fact shows that the Stein's method is robust. If $p \doteqdot 0.2$ (or 0.8), then the Stein's is robust for the mixture of order $\Delta < 3$. The same is true for order $\Delta < 2.5$, $\Delta < 2$ and $\Delta < 1.5$ in case of $p \doteqdot 0.3$ (0.7), $p \doteqdot 0.4$ (0.6) and $p \doteqdot 0.5$, respectively. Therefore, for the mixed distribution with parameter values which fall within these intervals, Eqs. (2) and (3) can be safely applied to the determination of sample size.

For the mixed distribution with parameter values which situate outside the above range, the Stein's is not so robust. Then it is recommendable to use Eqs. (16) and (17) instead of Eqs. (2) and (3) for sample size determination.

Table 2 The ratios of sample size expectation

	p	0.1	0.2	0.3	0.4	0.5
	Δ	0.9	0.8	0.7	0.6	
N_1 = 10	0.5	1.000	1.000	1.000	1.000	1.000
	1.0	0.999	1.000	1.002	1.004	1.004
	1.5	0.996	1.001	1.009	1.015	1.017
	2.0	0.991	1.004	1.023	1.037	1.042
	2.5	0.987	1.013	1.045	1.070	1.079
	3.0	0.988	1.026	1.074	1.116	1.123
	3.5	0.993	1.045	1.107	1.154	1.171
	4.0	1.004	1.067	1.143	1.199	1.220
	4.5	1.020	1.091	1.179	1.244	1.268
	5.0	1.038	1.116	1.214	1.287	1.314
N_1 = 20	0.5	1.000	1.000	1.000	1.001	1.001
	1.0	0.996	1.000	1.001	1.009	1.011
	1.5	0.986	1.002	1.021	1.036	1.041
	2.0	0.971	1.009	1.053	1.083	1.093
	2.5	0.954	1.025	1.098	1.146	1.162
	3.0	0.942	1.049	1.152	1.218	1.241
	3.5	0.936	1.080	1.211	1.295	1.324
	4.0	0.937	1.114	1.273	1.372	1.406
	4.5	0.942	1.151	1.334	1.448	1.486
	5.0	0.951	1.188	1.393	1.520	1.563
N_1 = 30	0.5	1.000	1.000	1.001	1.001	1.001
	1.0	0.995	1.000	1.007	1.013	1.015
	1.5	0.981	1.003	1.030	1.048	1.055
	2.0	0.961	1.016	1.071	1.108	1.121
	2.5	0.942	1.040	1.130	1.188	1.207
	3.0	0.928	1.074	1.199	1.277	1.303
	3.5	0.922	1.116	1.274	1.371	1.403
	4.0	0.923	1.161	1.350	1.465	1.503
	4.5	0.930	1.208	1.426	1.557	1.601
	5.0	0.940	1.256	1.499	1.646	1.694

5. Concluding Remarks

Although this paper postulates that the prior informatibns abot p and μ_2 are known, accurate prior informations are often unusual. Therefore the dependency of sample size on the prior informations is discussed and the robustness of Stein's method for mixed normal distribution is referred from such a situation.

The method suggested here is also applicable to the determination of sample size when a nuisance factor comes into the stable system. An example is perceived in estimation of the mean production rate or the mean production time in a stable production line participated by non-skilled person. Since the prior informations can be obtained to some extent even in this situation, the assumptions made in this paper are not so unreasonable.

References

1) J. Behboodian, Technometrics, **14**, 919 (1972).
2) M. G. Kendall and A. Stuart, The Advanced Theory of Statistics Vol. 1, 3rd ed., p.231, Griffin, London (1969).
3) G. Nadler, Motion and Time Study, p. 371, McGraw-Hill, New York (1955).
4) C. Stein, Ann. Math. Stat., **16**, 243 (1945).