A Sample Size Determination in the Non-normal Population

A Sample Size Determination in the Non-normal Population

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

(Received November 10, 1983)

The sample size determination is keenly required in statistical experiments and its related fields. Stein's two-stage procedure has been applied in many opportunities which assumes that the population is normally distributed. Many actual observations taken from a non-normal population do not necessarily yield appropriate sample sizes that are suggested by Stein's method. This paper, based on Stein's procedure, proposes a sample size determination method applicable to a non-normal circumstance. The detailed discussions are illustrated as to two cases: i) the population is non-normal, and ii) the observations have correlations between themselves.

1. Introduction

Stein's two-stage procedure[7] including its improvements, *e.g.*, a method using sample range[4], is frequently used for determination of sample size in statistical surveys. All these procedures postulate a normal population as the underlying distribution. In actual situation, however, the populations from which the samples are drawn may not warrant this postulation. Consequently Stein's method can not always determine an appropriate sample size if the population distribution shows a non-normal pattern. In order to pull through such a situation, it is necessary to develop an improved method applicable even to the non-normal populations.

Gayen[2] and Geary[3] investigate the distribution of Student's t-statistic in non-normal population, and derive the approximate distributions of t-statistic. Srivastava[6] studies the effect of non-normality on the power function of t-test, and obtains the expression of the power function of Student's t-test. Discussing the power function of Stein's two-sample method, Bhattachrjee[1] insists that Stein's method fails to attain the expected confidence coefficient. He does not give any procedure suitable to non-normal populations.

This paper proposes a modified sample size determination method, based on Stein's work, in the following two cases:

i) the population distribution is not normal and

ii) a correlation between observations is perceived.

2. Brief outline of Stein's two-sample scheme

Stein[7] showed that two-stage sampling scheme must be utilized to estimate a mean of normal population $N(\mu, \sigma^2)$ when we assign an appropriate length to a confidence interval of estimate *a priori*. His procedure is summarized briefly as follows:

* Ryukoku University, Faculty of Business Administration.

** Department of Industrial Engineering, College of Engineering.

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 total sample size N by

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

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.

3. Construction of a procedure for non-normal population

3.1 Distribution of t-statistic

If the population distribution is not normal, the Student's t-statistic, $t=(\bar{x}-\mu)/(s/\sqrt{n})$, does no longer be distributed in Student manner, where $\bar{x}=\sum_{i=1}^{n} x_i/n$, $s^2=\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)$. This fact suggests that the α-point $t'(\alpha)$ of t-statistic becomes distinct from the ordinary $t_{n-1}(\alpha)$ which appears in Eq. (1) in the normal case.

Gayen[2] deduces the t-statistic distribution for a non-normal population by assuming that its density is specified by the first four terms of an Edgeworth series

$$f(x)=\phi(x)-(\lambda_3/3!)\phi^{(3)}(x)+(\lambda_4/4!)\phi^{(4)}(x)+(10\lambda_3^2/6!)\phi^{(6)}(x)\,,$$

where $\lambda_3=\kappa_3/\kappa_2^{3/2}$, $\lambda_4=\kappa_4/\kappa_2^2$, κ_i denotes the i-th order cumulants, $\phi(x)$ the density of standard normal distribution and $\phi^{(r)}(x)=(d/dx)^{(r)}\phi(x)$. As a result, he indicates that the lower probability, $P(t_0)=Pr\{t\leqq -t_0\}$ of t, and the upper, $P'(t_0)=Pr\{t\geqq t_0\}$, are given by

$$P(t_0)=P_0(t_0)+\lambda_3 P_{\lambda_3}(t_0)-\lambda_4 P_{\lambda_4}(t_0)+\lambda_3^2 P_{\lambda_3^2}(t_0)\,, \tag{2}$$

and

$$P'(t_0)=P_0(t_0)-\lambda_3 P_{\lambda_3}(t_0)-\lambda_4 P_{\lambda_4}(t_0)+\lambda_3^2 P_{\lambda_3^2}(t_0)\,, \tag{3}$$

respectively. Each term in Eqs. (2) and (3) is defined as

$$P_0(t_0)=I_{u_0}(\nu/2,\ 1/2)/2\,,$$

$$P_{\lambda_3}(t_0)=\{1+(2\nu+1)t_0^2/\nu\}/\{6\sqrt{2\pi(\nu+1)}\,T_0^{(\nu+2)/2}\}\,,$$

$$P_{\lambda_4}(t_0)=(\nu/24)I_{u_0}(\nu/2,\ 1/2)-\{\nu(\nu+3)/(12(\nu+1))\}\cdot I_{u_0}((\nu+2)/2,\ 1/2)+\{\nu(\nu+5)/(24(\nu+1))\}I_{u_0}((\nu+4)/2,\ 1/2)\,,$$

$$\begin{aligned}P_{\lambda_3^2}(t_0)=&\{\nu(2\nu+7)/72\}I_{u_0}(\nu/2,\ 1/2)\\&-\{\nu(2\nu^2+9\nu+15)/(24(\nu+1))\}I_{u_0}((\nu+2)/2,\ 1/2)\\&+\{\nu(2\nu^2+9\nu+19)/(72(\nu+1))\}\{3I_{u_0}((\nu+4)/2,\ 1/2)\\&-I_{u_0}((\nu+6)/2,\ 1/2)\}\,,\end{aligned}$$

where $T_0=1+t_0^2/\nu=u_0^{-1}$ and $\nu=n-1$, and $I_{u_0}(a, b)$ denotes the incomplete beta function

$$I_{u_0}(a, b)=\int_0^{u_0} u^{a-1}(1-u)^{b-1}du \;/\; \int_0^1 u^{a-1}(1-u)^{b-1}du\,.$$

Eqs. (2) and (3) enable us to calculate the desired α-point $t'(\alpha)$ of the t-statistic distribution for a non-normal population.

3.2 Modification of sample size in a non-normal distribution

Substitution of $\nu=N_1-1$ into Eqs. (2) and (3) gives t_0 and t_0' which satisfy

$$P(t_0)=P'(t_0')=\alpha/2\,.$$

Let us denote t_0 and t_0' as $-L(\alpha/2)$ and $U(\alpha/2)$, respectively, then

$$Pr\{L(\alpha/2)\leqq(\bar{x}-\mu)/(s/\sqrt{N})\leqq U(\alpha/2)\}=1-\alpha\,.$$

Equating the length of confidence interval of $\bar{x}$ to the preassigned value $2l$ in the above formula, we obtain

$$U(\alpha/2)s/\sqrt{N}-L(\alpha/2)s/\sqrt{N}=2l\,.$$

Replacement of s by an estimate $\hat{\sigma}$ in the first sample leads to a formula for determining the modified sample size as

$$N=[\{U(\alpha/2)-L(\alpha/2)\}/\{2t_{N_1-1}(\alpha)\}]^2[t_{N_1-1}(\alpha)\hat{\sigma}/l]^2\,. \qquad (4)$$

The second term of the product in Eq. (4) is equivalent to Stein's one shown in Eq. (1), and the first term implies the proposed coefficient of modification:

$$C=[\{U(\alpha/2)-L(\alpha/2)\}/\{2t_{N_1-1}(\alpha)\}]^2\,.$$

Table 1 shows the coefficient of modification for several values of parameters λ_3, λ_4, α, and N_1. If the population is normal, it follows from $\lambda_3=\lambda_4=0$ that the coefficient of modification C becomes equal to unity. In the case of normal distribution, therefore, Eq. (4) is reduced to Stein's formula.

As is seen in Table 1, the values of coefficient of modification are symmetrical about λ_3 with respect to its origins. Generally the greater the absolute value of λ_3 is, the more skewed the distribution becomes. This property inclines to give a sample size larger than Stein's. As regards λ_4 in Table 1, little functional relationship between λ_4 and the coefficient of modification is perceived. In other words, any relation between λ_4 and the sample size can not be detected. This tendency may occur as the result of compounded effect of λ_4, the first sample size N_1, and confidence coefficient $1-\alpha$.

4. Modification of sample size in correlated observations

In what follows, the case where the Stein's assumption concerning independency in the population is not justified will be discussed. Since it is so difficult to

treat generally the problem, let us bring it to a simplified case in which the correlations between only two adjacent observations exist. This assumption is actual, because most consecutive values are subject to the correlation in observation work. And assume that

$$E(x_i)=\mu, \quad V(x_i)=\sigma^2,$$

Table 1 Coefficient of modification, C

$N_1=10$ $\alpha=0.1$

λ_4 \ $\lvert\lambda_3\rvert$	0	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0
−3	1.042	1.054	1.089	1.145	1.216	1.296	1.376	1.450	1.519	1.580	1.637
−2	1.027	1.038	1.073	1.129	1.202	1.285	1.369	1.448	1.519	1.583	1.641
−1	1.013	1.024	1.058	1.114	1.188	1.274	1.362	1.444	1.519	1.586	1.645
0	1.000	1.011	1.044	1.099	1.174	1.262	1.353	1.440	1.519	1.588	1.649
1	0.988	0.999	1.031	1.085	1.161	1.250	1.345	1.435	1.517	1.590	1.653
2	0.978	0.988	1.019	1.072	1.147	1.238	1.336	1.430	1.516	1.591	1.657
3	0.968	0.977	1.007	1.059	1.133	1.226	1.326	1.424	1.513	1.592	1.660
4	0.958	0.968	0.996	1.047	1.120	1.213	1.315	1.417	1.510	1.592	1.663
5	0.950	0.959	0.986	1.035	1.107	1.200	1.305	1.410	1.507	1.592	1.666
6	0.942	0.950	0.977	1.024	1.095	1.187	1.294	1.402	1.502	1.591	1.668

$N_1=10$ $\alpha=0.2$

λ_4 \ $\lvert\lambda_3\rvert$	0	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0
−3	0.973	0.981	1.005	1.049	1.119	1.216	1.338	1.472	1.607	1.736	1.856
−2	0.983	0.990	1.013	1.055	1.121	1.215	1.334	1.468	1.604	1.734	1.856
−1	0.992	0.999	1.020	1.060	1.123	1.214	1.330	1.463	1.600	1.733	1.856
0	1.000	1.007	1.027	1.065	1.126	1.212	1.326	1.459	1.597	1.731	1.856
1	1.008	1.014	1.034	1.071	1.128	1.212	1.323	1.455	1.594	1.729	1.856
2	1.016	1.022	1.041	1.076	1.130	1.211	1.320	1.450	1.590	1.727	1.856
3	1.023	1.029	1.047	1.080	1.133	1.211	1.316	1.446	1.586	1.725	1.856
4	1.030	1.035	1.053	1.085	1.135	1.210	1.313	1.441	1.582	1.723	1.855
5	1.036	1.042	1.059	1.089	1.138	1.210	1.310	1.437	1.578	1.720	1.855
6	1.043	1.048	1.064	1.094	1.141	1.210	1.308	1.432	1.574	1.717	1.854

$N_1=20$ $\alpha=0.1$

λ_4 \ $\lvert\lambda_3\rvert$	0	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0
−3	1.007	1.013	1.032	1.063	1.107	1.163	1.226	1.292	1.357	1.416	1.470
−2	1.004	1.010	1.029	1.060	1.104	1.159	1.223	1.290	1.355	1.417	1.472
−1	1.002	1.008	1.026	1.056	1.100	1.155	1.219	1.287	1.354	1.416	1.473
0	1.000	1.006	1.023	1.053	1.096	1.151	1.215	1.284	1.352	1.416	1.474
1	0.998	1.004	1.021	1.050	1.092	1.147	1.211	1.281	1.351	1.416	1.475
2	0.996	1.001	1.018	1.047	1.088	1.143	1.207	1.278	1.349	1.415	1.476
3	0.994	1.000	1.016	1.044	1.085	1.139	1.203	1.275	1.346	1.415	1.476
4	0.992	0.998	1.014	1.041	1.081	1.135	1.199	1.271	1.344	1.414	1.477
5	0.991	0.996	1.011	1.038	1.078	1.131	1.195	1.268	1.342	1.413	1.477
6	0.989	0.994	1.009	1.036	1.075	1.127	1.191	1.264	1.339	1.411	1.477

Table 1 (Continued)

$N_1=20$ $\alpha=0.2$

λ_4 \ $\lvert\lambda_3\rvert$	0	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0
−3	0.976	0.980	0.991	1.011	1.041	1.084	1.143	1.221	1.316	1.422	1.530
−2	0.984	0.988	0.999	1.018	1.047	1.089	1.147	1.223	1.316	1.421	1.530
−1	0.992	0.996	1.006	1.025	1.054	1.094	1.151	1.225	1.317	1.421	1.529
0	1.000	1.003	1.014	1.032	1.060	1.100	1.155	1.227	1.318	1.421	1.529
1	1.008	1.011	1.021	1.039	1.066	1.105	1.158	1.229	1.318	1.420	1.529
2	1.015	1.018	1.028	1.045	1.072	1.110	1.162	1.232	1.319	1.420	1.528
3	1.022	1.025	1.035	1.052	1.078	1.115	1.166	1.234	1.320	1.420	1.528
4	1.029	1.032	1.042	1.058	1.084	1.120	1.170	1.236	1.320	1.420	1.527
5	1.036	1.039	1.048	1.065	1.089	1.125	1.173	1.239	1.321	1.420	1.527
6	1.043	1.046	1.055	1.071	1.095	1.130	1.177	1.241	1.322	1.420	1.527

$N_1=30$ $\alpha=0.1$

λ_4 \ $\lvert\lambda_3\rvert$	0	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0
−3	1.001	1.005	1.018	1.039	1.069	1.109	1.157	1.211	1.267	1.324	1.377
−2	1.001	1.005	1.017	1.038	1.068	1.107	1.155	1.209	1.266	1.323	1.377
−1	1.001	1.004	1.016	1.037	1.067	1.105	1.153	1.207	1.264	1.322	1.377
0	1.000	1.004	1.016	1.036	1.065	1.104	1.151	1.205	1.263	1.321	1.377
1	1.000	1.003	1.015	1.035	1.064	1.102	1.149	1.203	1.261	1.320	1.377
2	0.999	1.003	1.014	1.034	1.062	1.100	1.147	1.201	1.259	1.319	1.376
3	0.999	1.003	1.014	1.033	1.061	1.098	1.145	1.199	1.258	1.318	1.376
4	0.999	1.002	1.013	1.032	1.060	1.097	1.143	1.197	1.256	1.316	1.375
5	0.998	1.002	1.013	1.031	1.059	1.095	1.141	1.195	1.254	1.315	1.375
6	0.998	1.001	1.012	1.030	1.057	1.093	1.139	1.192	1.252	1.314	1.374

$N_1=30$ $\alpha=0.2$

λ_4 \ $\lvert\lambda_3\rvert$	0	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0
−3	0.982	0.984	0.992	1.004	1.023	1.049	1.084	1.131	1.190	1.263	1.348
−2	0.988	0.990	0.998	1.010	1.028	1.054	1.089	1.134	1.193	1.265	1.349
−1	0.994	0.996	1.003	1.016	1.034	1.059	1.093	1.138	1.195	1.266	1.349
0	1.000	1.002	1.009	1.021	1.039	1.064	1.097	1.141	1.198	1.268	1.350
1	1.006	1.008	1.015	1.027	1.044	1.069	1.102	1.145	1.200	1.269	1.351
2	1.012	1.014	1.021	1.032	1.050	1.074	1.106	1.148	1.203	1.271	1.351
3	1.017	1.020	1.026	1.038	1.055	1.078	1.110	1.152	1.206	1.273	1.352
4	1.023	1.025	1.032	1.043	1.060	1.083	1.114	1.155	1.208	1.274	1.353
5	1.028	1.031	1.037	1.048	1.065	1.088	1.118	1.159	1.211	1.276	1.354
6	1.034	1.036	1.042	1.053	1.070	1.092	1.123	1.162	1.214	1.280	1.355

$$cov(x_i, x_j) = \begin{cases} \rho\sigma^2, & (|i-j|=1) \\ 0, & (|i-j| \geqq 2) \end{cases}$$

and autocorrelation coefficient ρ ranges from −0.5 to 0.5. Although the interval of ρ is theoretically between −1 to 1, practical ρ is almost covered by the range between −0.5 to 0.5[5]. In consequence with no loss in generality, one may proceed

to the discussion on such an interval.

From the above assumptions

$$E(\bar{x})=\mu, \quad E(\hat{\sigma}^2)=\sigma^2(1-2\rho/n),$$
$$V(\bar{x})=(\sigma^2/n)\{1+2\rho(1-1/n)\}$$
$$=(\sigma^2/n)(1+2\rho)+0(1/n^2),$$

which suggest that the expected value of $\hat{\sigma}^2$ tends to σ^2 as n becomes larger, and the effect of ρ on the variance of $\bar{x}$ is important.

Stein's derivation gives the modified sample size equation such as

$$N=(1+2\rho)[t_{N_1-1}(\alpha)\hat{\sigma}/l]^2. \tag{5}$$

In the extreme case, such as $\rho=-0.5$, N from Eq. (5) becomes zero, that is, the variance of $\bar{x}$ and the length of confidence interval become zero simultaneously. Conversely, if $\rho=0.5$, a sample size two times as large is necessary than in the case of independent observations.

Provided that a correlation exists between the observations themselves, the sample size in Stein's method becomes over-weighted toward the negative ρ and under-emphasized with the positive ρ. The appropriate sample size is given by Eq. (5) with the correction factor $(1+2\rho)$.

5. Illustrative examples

A simulation study certifies the efficiency of the proposed sample size modifi-

Table 2 The results of simulated modification of sample size (Non-normal population)

Population Distribution	Number of $\bar{x}$ out of conf. int.	
	Stein's	Modified
Doubly exponential (Min)	129	111
$\lambda_3=-1.1$	120	100
$\lambda_4=2.4$	141	118
$C=1.1248$	111	94
Exponential	148	127
$\lambda_3=2$	148	116
$\lambda_4=6$	158	101
$C=1.3661$	184	124
Gamma	102	99
$\lambda_3=0.5$	106	104
$\lambda_4=0.375$	117	109
$C=1.0254$	122	113
Uniform	91	87
$\lambda_3=0$	100	98
$\lambda_4=-1.2$	98	102
$C=1.0154$	88	92

cation method as is explained below. Each parameter is given as $\mu=4000$, $\sigma^2=1000^2$, $2l=400$, $1-\alpha=0.9$, $N_1=10$. The probability of $\bar{x}$ lying outside the interval $\mu-l=3800\leqq\bar{x}\leqq\mu+l=4200$ is expected to be approximately ten percent.

Table 2 shows the effects of the modification for four different non-normal populations, and the results by Stein's method are also shown for comparative purposes. Table 3 summarizes the effects of the methods for correlated observations. It is evident from Tables 2 and 3 that the proposed method gives the appropriate sample size for non-normal population and correlated observations. Whereas in the case of original Stein's method, it seems that the condition $\alpha=0.1$ is unsatisfied.

Table 3 The results of simulated modification of sample size (Observations with correlation)

ρ	Number of $\bar{x}$ out of conf. int.	
	Stein's	Modified
0.5	239	123
0.4	209	112
0.2	181	118
0.0	108	108
−0.2	35	90
−0.4	2	108
−0.5	0	38

6. Concluding remarks

If the population distribution shows non-normal pattern, the application of Stein's original method to determination of sample size leads to unsatisfactory results. In such a case, multipling a correcting factor brings the method valid with the assigned conditions satisfactory. The correcting factor is given in the table of coefficients. The modification of sample size for correlated observations is also discussed. The modifying coefficient for them is given in a simple formula.

References

1) G.P. Bhattacharjee, Ann. Math. Stat., **36**, 651 (1965).
2) A.K. Gayen, Biometrika, **36**, 353 (1949).
3) R.C. Geary, ibid., **34**, 209 (1947).
4) G. Nadler, Motion and Time Study, 371, McGraw-Hill, New York, 1955.
5) H. Scheffe, The Analysis of Variance, John Wily & Sons, New York, 1959.
6) A.B.L. Srivastava, Biometrika, **45**, 421 (1958).
7) C. Stein, Ann. Math. Stat., **16**, 243 (1945).