자료의 범위를 이용한 표본의 크기결정

Using the sample Range as a basis for Calculating Sample Size

한 근 식*
Geunshik Han

I. 서 론

표본설계에서 중요한 문제중의 하나가 표본의 크기결정이다. 표본의 크기 n에 따라서 조사비용, 검정력(power)등에 차이가 있게 마련이다. 우선 허용오차에 의한 표본의 크기결정식(Cochran ,1977)을 살펴보자.

표본조사결과 모수 θ와 추정치 $\hat{\theta}$은 차이가 있게 마련이다. 이때 최대한 허용할 수 있는 차이를 허용오차(B라 하자)라 하며 허용오차를 초과할 확률이 α가 되도록 표본의 크기를 결정하고자 한다. 모평균 $\overline{Y}$의 추정시 허용오차를 B라 하고 이를 확률로 표현하면 다음과 같다.

$$P(|\overline{Y}-\overline{y}| \geq B) = \alpha \quad \text{----------(1)}$$

이때 허용오차 B와 $\overline{y}$의 표준오차간에는 다음과 같은 식이 성립한다.

$$B = t_{\frac{\alpha}{2}} \sqrt{V(\overline{y})} \quad \text{----------(2)}$$

위식에서 $t_{\frac{\alpha}{2}}$는 t분포에서 $\Pr(|t| \geq t_{\frac{\alpha}{2}}) = \alpha$ 를 만족하는 상수이다. $V(\overline{y})$는 표본의 크기 n의 함수이므로 오차의 한계 B가 주어졌을 때 다음 관계식

$$t_{\frac{\alpha}{2}} \sqrt{V(\overline{y})} = t_{\frac{\alpha}{2}} \sqrt{\frac{S^2}{n} \frac{N-n}{N}} = B$$

를 표본의 크기 n에 대해서 정리하면, 모평균 $\overline{Y}$의 추정오차의 한계가 B가 되도록하는 표본의 크기 n은 다음과 같다.

* 한신대 정보시스템공학과

$$n = \frac{NS^2}{ND + S^2} \quad \text{----------(3)}$$

여기에서 $D = \left(\frac{B}{t_{\frac{\alpha}{2}}}\right)^2$.

95%신뢰구간의 경우 표본의 크기가 상당히 크다면 $t_{\frac{\alpha}{2}} = 1.96$ 을 이용하며 일반적으로 추정량 $\bar{y}$에 대한 표준편차의 2배가 되도록 추정오차의 한계를 설정하므로 실제표본 크기 결정시 $t_{\frac{\alpha}{2}} = 2$ 를 이용한다.

만약 모집단의 크기 $N \rightarrow \infty$이면 위의 식(3)은 다음과 같다.

$$n = \frac{(tS)^2}{B^2} \quad \text{----------(4)}$$

II. S^2 (σ^2) 의 추정

위의 식(3)과 (4)에서 보는바와같이 표본의 크기 n의 결정식은 표본조사를 통해 추정하려는 S^2 (σ^2)에 대한 정보가 요구된다. 결국 표본의 크기 n을 추정하기 위해 S^2의 추정치를 구해야 하는데 몇가지 널리 알려진 방법을 방법을 소개하기로 한다.

1. 사전에 S가 알려져 있다. (거의 현실성이 없다.)

2. 특정조사가 매년 또는 격년으로 시행되는 경우 전년도 혹은 직전의 조사에서 추정된 S 혹은 CV값을 이용한다.

3. 조사 연구자 A와 같은 연구 목적으로, 같은 혹은 유사한 모집단에 대한 조사가 이루어진 경우, 조사 연구자 B는 A의 조사 연구 결과로서 추정된 S를 이용할 수 있다. 이와 같은 상황은 흔히 볼 수 있는 것으로 대통령 선거에 앞서 많은 언론기관 혹은 여론 기관에서 표본조사를 시행한 결과에서 추정된 S를 이용하여 표본의 크기를 결정할 수 있다.

4. 자료의 범위(range)를 이용할 수 있다. 측정된 자료가 정규 분포를 따르는 경우, 즉 크기가 n인 표본이 정규분포에서 추출되었을 때 자료의 범위 (R)와 표준편차 (S)의 비 (ratio), $\frac{R}{S}$를 Tippett은 <표 1> 과 같이 정리하였다.

Tippett의 표를 이용하는 방법을 다음의 예를 이용하여 설명하겠다.

경기 남부지역의 각 주유소에서 판매되는 가솔린의 1리터 당 평균을 95%신뢰구간에서 추정오차의 한계가 5원이 되도록 추정하려 한다. 조사 연구자가 알아본 결과 가장 싼 가격을 995원, 가장 비싼 가격을 1190원이었다. 이 정보에 의하면 가격의 범위는 R=195원이 된다.

이제 <표.1>을 활용하기 위해 예상되는 표본의 크기를 100이라 하자.

<표.1>에서 n = 100일 때 $\frac{R}{S}$ = 5.03이다.

R = 195를 이용하면 표준 편차는 S = 38.76원으로 추정된다.

이제 S=38.76을 식(4.7)에 대입하면

$$n=\frac{(1.96*38.76)^2}{5^2}=230.8$$

을 얻는다.

다시 <표.1>에서 표본의 크기, n=231에 해당하는 $\frac{R}{S} \approx 5.57$이다.

R = 195를 이용하면 표준 편차는 S = 35원으로 추정된다.

이제 $S = 35$를 식(4.7)에 대입하면

$$n=\frac{(1.96*35)^2}{5^2}= 188.2$$

를 얻는다.

다시 <표.1>로 돌아가서 n=188일때 $\frac{R}{S} \approx 5.45$를 구한다.

R = 195를 이용하면 표준 편차는 S = 35.78원으로 추정된다.

이제 S = 35.78를 식(4.7)에 대입하면

$$n=\frac{(1.96*35.64)^2}{5^2}=196$$

을 얻는다.

이와 같은 절차를 반복하면 n의 값은 196, 194, 195, 194 등으로 계산되며 최종적으로 n이 194에 수렴하게 된다. 결국 95% 신뢰구간에서 오차의 한계가 5원이 되도록 추정하려면 194개의 주유소를 조사해야 한다.

n	$\frac{R}{S}$	n	$\frac{R}{S}$	n	$\frac{R}{S}$
2	1.13	22	3.83	175	5.40
3	1.69	23	3.86	200	5.49
4	2.06	24	3.90	225	5.57
5	2.33	25	3.93	250	5.64
6	2.53	26	3.96	275	5.70
7	2.60	27	4.00	300	5.76
8	2.85	28	4.03	350	5.85
9	2.97	29	40.6	400	5.94
10	3.08	30	40.9	450	6.01
11	3.17	35	4021	500	6.07
12	3.26	40	4.32	550	6.13
13	3.34	45	4.42	600	6.18
14	3.41	50	4.50	650	6.23
15	3.47	60	4.64	700	6.27
16	3.53	70	4.75	750	6.32
17	3.59	80	4.85	800	6.35
18	3.64	90	4.94	850	6.39
19	3.69	100	5.03	900	6.42
20	3.74	125	5.17	950	6.45
21	3.78	150	5.30	1000	6.48

<표. 1> Tippett의 $\frac{R}{S}$

III. 자료의 범위를 이용한 σ의 추정

M_1 : $\sigma \approx \frac{R}{6}$ (1979, Daniel 과 Terrell)

M_2 : $\sigma \approx \frac{R}{4}$ (1971, Mendenhall, Ott, Scheaffer)

M_1은 모집단이 정규분포를 따를 때,

$$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$$

이라는 경험적 사실에 근거를 둔 것으로 $\frac{R}{6}$은 σ의 적절한 추정치를 제공한다.

M_2은 모집단이 정규분포를 따를 때,

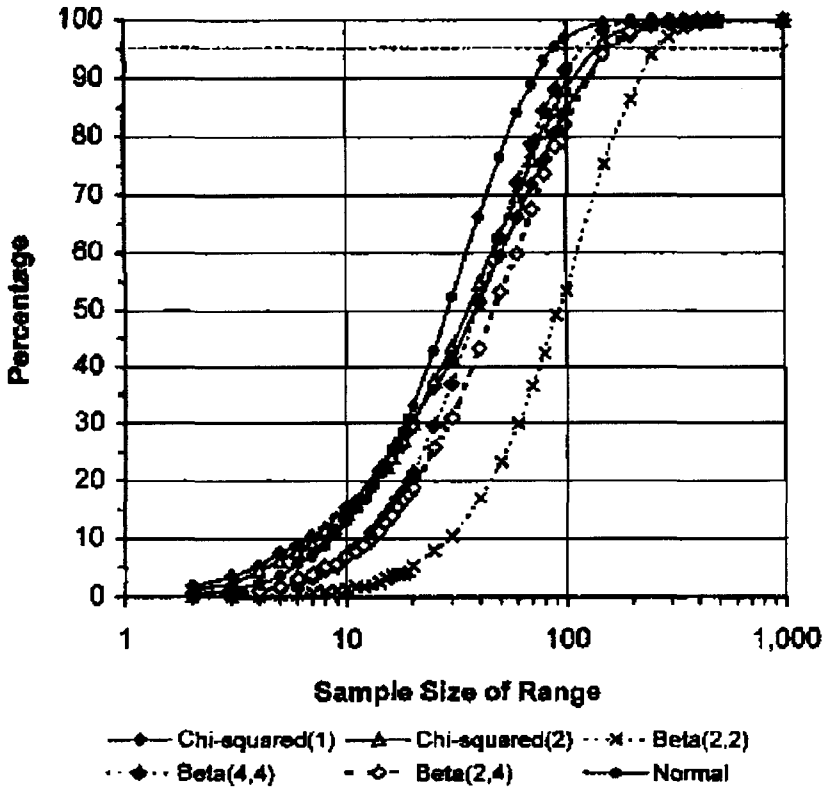
$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.954$$

이라는 경험적 사실에 근거를 둔 것으로 $\frac{R}{4}$ 역시 σ의 적절한 추정치를 제공한다

표본의 크기 결정식(4)에 의하면 M_1, M_2에 의해 추정된 σ의 값이 실제 σ보다 크다면 표본의 크기 n은 불필요하게 커지게 되나 검정력(power)은 충분히 보장된다. 반면에 추정된 σ가 실제 σ보다 작다면 표본의 크기 n은 작아지나 검정력(power)은 떨어진다. 결국 표본설계시 요구되는 검정력을 보장하기 위해서는 추정된 σ의 값이 실제 σ보다 크거나 같게 추정되어야한다. 그러나 위에 언급한 두 방법 M_1, M_2에서는 $\hat{\sigma} \geq \sigma$ 지 아닌지를 알 수가 없다.

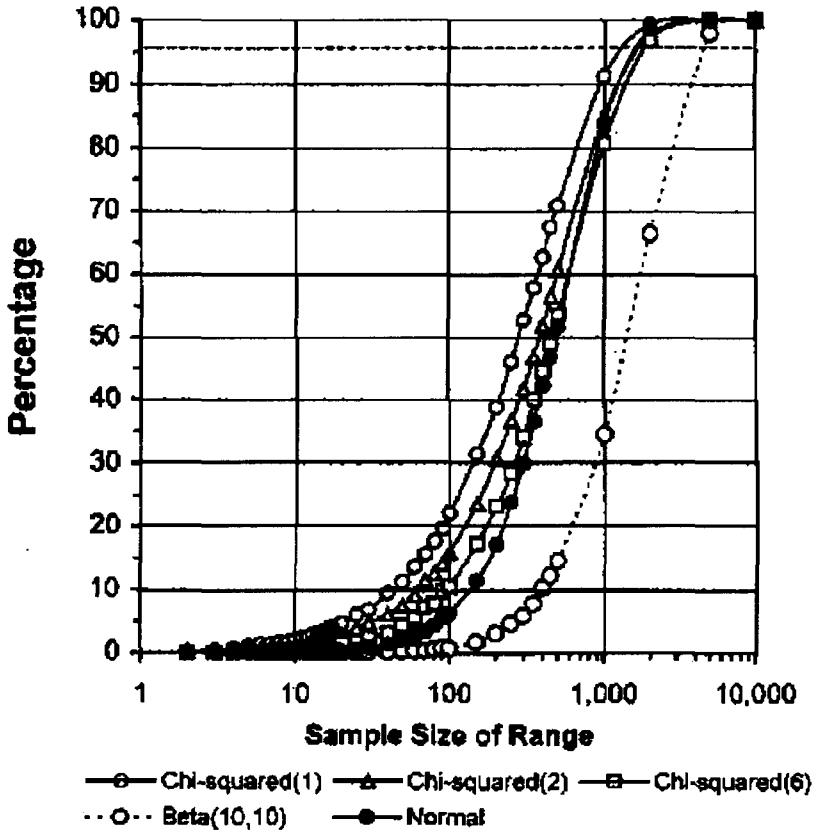
IV. Browne의 power - preserving estimator

Browne는 Chi-squqre(1), Chi-squqre(2), Beta(2, 2), Beta(4, 4), Beta(2, 4), 정규분포등을 이용하여 크기가 n인 10,000개의 독립표본을 발생시켜 $K = \frac{R}{\sigma}$의 값을 계산하여 $\hat{\sigma} \geq \sigma$ 가능성과 표본의 크기와의 관계를 연구하였다.



위 그림은 $\frac{R}{4}$에 의해 σ를 추정하는 경우, $\frac{R}{4}$이 σ보다 크거나 같게되는 가능성을 보여주고 있다. 위 그림에 의하면 표본의 크기 n이 30보다 작은 경우 $\frac{R}{4}$에 의해 추정된 σ가 실제 σ보다 크거나 같을 가능성이 50%가 되지 않는다는 것을 알 수 있다. 정규분포의 경우 n=87보다 크거나 같게될 때 $\frac{R}{4}$에 의해 추정된 σ가 실제 σ보다 크거나 같을 가능성이 95%에 달한다. 표본의 크기가 263보다 클때에야, $\frac{R}{4}$에 의해 추정된 σ가 실제 σ보다 크거나 같을 가능성이 모든 분포에 대해서 95%에 달한다.

다음 그림은 $\frac{R}{6}$를 이용하여 σ를 추정할때의 표본의 크기와 $\frac{R}{6} \geq \sigma$가능성을 보여주고 있다. 이 그림에 의하면 n=1,000일 때 $\frac{R}{6}$는 언급한 모든 분포에 대해서 검정력을 보장하지 못하고 있다. 정규분포의 경우 n=1750일 때 검정력을 보장하고 있다.



아래 표는 검정력을 보장해주는, 즉 $\frac{R}{K} \geq \sigma$ 될 가능성이 적어도 95%가되는 K값을 5개 분포와 표본의 크기에 따라 작성한 것이다.

표본의 크기	정규분포	$\chi^2(1)$	B(0.5, 0.5)	Uniform	B(2, 2)
5	1.04	0.63	1.15	1.17	1.14
10	1.86	1.27	2.16	2.07	2.03
20	2.62	1.93	2.61	2.71	2.72
30	3.04	2.33	2.72	2.95	3.04
40	3.30	2.59	2.77	3.07	3.23
50	3.50	2.82	2.79	3.14	3.37
60	3.68	2.99	2.80	3.20	3.47
70	3.79	3.17	2.81	3.24	3.54
80	3.91	3.29	2.81	3.26	3.60
90	4.03	3.42	2.82	3.28	3.66
100	4.12	3.52	2.83	3.30	3.70
250	4.81	4.41	2.83	3.40	3.99
500	5.29	5.12	2.83	3.43	4.13
1,000	5.75	5.80	2.83	3.45	4.23
10,000	7.09	8.06	2.83	3.45	4.40

<표.2> $\hat{\sigma}$ 이 σ 보다 크거나 같게될 가능성이 95%이상이 되도록하는 K값

<참 고 문 헌>

Browne, R. H.(2001) Using the Sample Range as a Basis for Calculating Sample Size in Power Calculations. The American Statistician, Vol. 55, No. 4 293-298

Cochran, W. G(1977) Sampling Techniques. 3rd edition, Wiley & Sons

Daniel, W. W. and Terrell, V. C.(1979) Business Statistics : Basic Concepts and methodology.

Mendenhall, W., Ott, L., and Scheaffer, R.L.(1971) Elementary Survey sampling Duxberry Press.

Tippett, L. H. C. (1925) On the extreme individuals and the range of samples taken from a normal population. Biometrika, Vol. 17 364-387