The Korean Communications
in Statistics Vol. 9, No. 3, 2002
pp. 877-888

Applications on p-values of Chi-Square Distribution

Chong Sun Hong[1], Sung Sick Hong[2]

Abstract

In this paper, behaviors and properties of p-values for goodness-of-fit test are investigated. With some findings on the p-values, we consider some applications to determine sample size of a survey research using the regression equation based on a pilot study data. Regression equations are obtained by the well-known least squared method, and we find that regression lines could be formulated with only two data points, alternatively. For further studies, this works might be extended to t distributions for testing hypotheses about population mean in order to determine sample size of a prospective study. Also similar arguments could be explored for F test statistics.

Keywords : Goodness-of-fit statistics, Least squared method, Pilot study, Sample size, Survey research.

1. Introduction

For data analysis of contingency tables, goodness-of-fit test statistics such as Pearson statistic X^2, Likelihood ratio statistic G^2, Freeman-Tukey statistic T^2, Power divergence statistic $I(\lambda)$ of Cressie and Read (1984), etc., are used frequently (see Bishop, Fienberg, and Holland (1975), Agresti (1990), and Christensen (1990)). Among many properties, all these goodness-of-fit test statistics which are followed Chi-square distribution have identical degrees of freedoms $(I-1)(J-1)$, even though sample sizes of $I\times J$ contingency tables are not equivalent.

When two dimensional contingency table, for example, is tested for H_0: two categorical variables are independent, one might encounter in practice not to get one's desired result which is the rejection of the null hypothesis. At this case, one seldom makes the count of cell doubled or tripled. Then each cell probabilities in a modified table are invariant, but results of analysis are easy to convert. Our purpose of this work is not to abuse these disadvantages.

1) Professor, Department of Statistics, SungKyunKwan University, Seoul, KOREA.
E-mail : cshong@skku.ac.kr
2) Concurrent Professor, Department of Internet Information Security, Hoseo Computer Technical College, Seoul, KOREA. E-mail : ogorch@hanmail.net

When one fails to reject the null hypothesis from the obtained contingency table, the previous survey might be regarded as a 'pilot study' of small sample size and let us suppose that a prospective survey research is demanded in near future. In order to obtain desired results, researchers might try to make the sample size of a prospective survey larger than that of the pilot study data analyzed previously. Under some assumptions and regularity conditions, cell probabilities of contingency table obtained from a pilot study are close to those of a prospective contingency table. Whereas goodness-of-fit test statistics for both contingency tables have different values, these statistics follow Chi-square distribution of the same degrees of freedom. Therefore, we need to research on p-values that utilize Chi-square distribution to decide an appropriate sample size of a prospective survey.

In Section 2, p-values of Chi-square distribution are defined and formulated when degrees of freedoms are even or odd numbers. And we will investigate properties of p-values of goodness-of fit test. With some findings of these properties, we may consider some applications to analyze contingency tables in Section 3. In other words, one make use findings to determine sample size of a survey research based on results of the pilot study. In Section 4, alternative behavior is developed to determine the sample size with ease. And further studies are discussed and provided in Section 5.

2. Definition and Calculation of p-values

2.1 Definition of p-value

Suppose a random variable X^2 follows Chi-square distribution with n ($n=1,2,3,...$) degrees of freedom, and let us define the p-value of the Chi-square distribution as the following $p(c)$:

$$p(c)\equiv \Pr(X^2>c)=1-\int_0^c \frac{1}{\Gamma(n/2)}\left(\frac{1}{2}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}\,dx\ , \tag{1}$$

where $c>0$. In order to get approximate values of (1), among others, the best known are the approximation of Fisher (1922) :

$$p(c)=1-\Phi(\sqrt{2c}-\sqrt{2n-1})$$

and the approximation of Wilson and Hilferty (1931) :

$$p(c) = 1 - \Phi(\{(c/n)^{\frac{1}{3}} - 1 + 2/9n\}\sqrt{9n/2}).$$

(See Kotz (1970) pp. 176-182 for more detail.) The exact p-value in (1) is formulated by some authors (eg, Park and Huh (1983) pp. 153). When the degree of freedom n is an even number, the p-value is obtained as

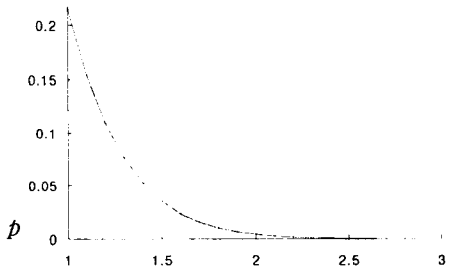
$$\begin{aligned} p(c) &= e^{-\frac{c}{2}} [\ 1 + \frac{(c/2)}{1!} + \frac{(c/2)^2}{2!} + \frac{(c/2)^3}{3!} + \cdots + \frac{(c/2)^{\frac{n}{2}-1}}{(n/2-1)!}\ ] \\ &= e^{-\frac{c}{2}} \sum_{i=0}^{\frac{n}{2}-1} [\ \frac{(c/2)^i}{i!}\ ] . \end{aligned} \tag{2}$$

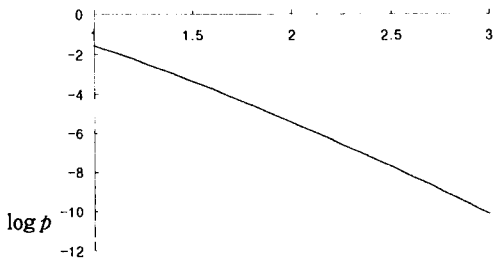
When n is an odd number ($n=2m+1,\ m=1,2,...$),

$$p(c) = 1 - \frac{1}{\sqrt{2\pi}} \int_0^c x^{-\frac{1}{2}} e^{-\frac{x}{2}} dx + \sqrt{2/\pi}\ e^{-\frac{c}{2}} \sum_{i=1}^{(n-1)/2} \frac{2^{i-1}\ (i-1)!\ c^{(2i-1)/2}}{(2i-1)!} . \tag{3}$$

2.2 Properties of $p(c)$

We can define $p(k\theta)$ as the probability of that the random variable of Chi-square distribution is greater than $k\theta$. In other words, the critical point c in (1) is denoted as $k\theta$, where $k \geq 1$ and $\theta > 0$. For any given n and θ, $p(k\theta)$ could be a function of k. These relationship between $p(k\theta)$ and k is represented at <Figure 1> for n=9 and θ=11.99, which are from an example of <Table 3>.





<Figure 1> Relationship between k and $p(k\theta)$ <Figure 2> Relationship between k and $\log p(k\theta)$

Now take the log of $p(k\theta)$, and data $\{(k, \log p(k\theta))\}$ are plotted at <Figure 2>. From <Figure 2>, we find that the relationship between k and $\log[p(k\theta)]$ is almost linear. With

this finding, one can estimate a simple regression line by using the least squared method. For the above example, a linear regression equation is fitted as

$$\log [\, \hat{p}(k\theta)] \;=\; 3.08 \;-\; 4.33\; k\;. \tag{4}$$

This estimated regression line (4) has some linearity information such that R^2 = 0.998 F-statistic value =7894.89 (p-value =0.0001).

3. Application I : Estimation of Sample Size for Survey Research

For a $I\times J$ two-dimensional contingency table in <Table 1> obtained from a pilot study, suppose that researchers fail to reach the desired result which is to reject H_0 : the two categorical variables are independent. Let us assume that a main survey research needs to be asked in near future. In order to obtain some desired results, researchers are willing to make sample size of the prospective survey larger than that of the pilot study. Then suppose that the sample size of the prospective survey research is k ($k\geq 1$) times of that of the pilot study without changing of any other situation of population for the prospective survey, where the data of the survey research is shown as a contingency table in <Table 2>. Each sample size of the pilot study and the prospective survey are N and kN, respectively. And x_{ij} and $x_{ij}^{(k)}$ are the corresponding cell values.

<Table 1> Pilot study data

	j=1	j=2	···	j=J	sum
i=1	x_{11}	x_{12}	···	x_{1J}	x_{1+}
i=2	x_{21}	x_{22}	···	x_{2J}	x_{2+}
···	···	···	···	···	···
i=I	x_{I1}	x_{I2}	···	x_{IJ}	x_{I+}
sum	x_{+1}	x_{+2}	···	x_{+J}	N

<Table 2> Prospective survey research data

	j=1	j=2	···	j=J	sum
i=1	$x_{11}^{(k)}$	$x_{12}^{(k)}$	···	$x_{1J}^{(k)}$	$x_{1+}^{(k)}$
i=2	$x_{21}^{(k)}$	$x_{22}^{(k)}$	···	$x_{2J}^{(k)}$	$x_{2+}^{(k)}$
···	···	···	···	···	···
i=I	$x_{I1}^{(k)}$	$x_{I2}^{(k)}$	···	$x_{IJ}^{(k)}$	$x_{I+}^{(k)}$
sum	$x_{+1}^{(k)}$	$x_{+2}^{(k)}$	···	$x_{+J}^{(k)}$	kN

Now assume that the characteristic of the population of a pilot study is invariant with that of a prospective survey, so that both (i,j)th cell probabilities, $\hat{p}_{ij}=x_{ij}/N$ of a pilot study and $\hat{p}_{ij}^{(k)}=x_{ij}^{(k)}/kN$ of a prospective survey converge to p_{ij} in probability. Then we obtain the following <Theorem>.

<Theorem>

If $\widehat{p_{ij}}$ of a pilot study converges to p_{ij} in probability, and $\widehat{p_{ij}}^{(k)}$ of prospective survey converges to p_{ij} in probability, then $\widehat{p_{ij}}^{(k)} - \hat{p}_{ij}$ converges to 0 in probability.

[Proof of Theorem]

Since $\{|\widehat{p_{ij}}^{(k)} - \hat{p}_{ij}| \geq \varepsilon\} \subset \{|\widehat{p_{ij}}^{(k)} - p_{ij}| \geq \varepsilon/2\} \cup \{|\widehat{p_{ij}} - p_{ij}| \geq \varepsilon/2\}$, the convergence of both $\widehat{p_{ij}}$ and $\widehat{p_{ij}}^{(k)}$ to p_{ij} in probability means that $\widehat{p_{ij}}^{(k)} - \hat{p}_{ij}$ converges to 0 in probability.

This <Theorem> says that if both $\widehat{p_{ij}}$ and $\widehat{p_{ij}}^{(k)}$ of a pilot study and a survey research, respectively, converge to the same p_{ij} in probability, then $\widehat{p_{ij}}$ of the pilot study are very much close to $\widehat{p_{ij}}^{(k)}$ of survey research, so that we find

$$\widehat{p_{ij}}^{(k)} = \widehat{p_{ij}} + o_p(1). \tag{5}$$

And one might assume that $x_{ij}^{(k)}$ in <Table 2> is very much close to the k times x_{ij} in <Table 1>, i.e.,

$$x_{ij}^{(k)} \simeq k \times x_{ij}\,. \tag{6}$$

For a pilot study, the Pearson Chi-square statistic X^2 is obtained :

$$X^2 = \sum_i \sum_j \frac{(x_{ij} - N\ \widehat{p_{ij}}^{0})^2}{N\ \widehat{p_{ij}}^{0}} \quad (= \theta, \text{ say}),$$

where $\widehat{p_{ij}}^{0} = \widehat{p_{i+}}^{0}\ \widehat{p_{+j}}^{0} = x_{i+}\,x_{+j}/N^2$, and where $x_{i+} = \sum_j x_{ij}$, $x_{+j} = \sum_i x_{ij}$, $\widehat{p_{i+}}^{0} = x_{i+}/N$, $\widehat{p_{+j}}^{0} = x_{+j}/N$. And for a prospective survey research, the Pearson Chi-square statistic $X^{2(k)}$ will be obtained :

$$X^{2(k)} = \sum_i \sum_j \frac{(x_{ij}^{(k)} - kN\ \widehat{p_{ij}}^{0(k)})^2}{kN\ \widehat{p_{ij}}^{0(k)}} \quad (= \theta^{(k)}, \text{ say}),$$

where $\widehat{p}_{ij}^{0(k)} = \widehat{p}_{i+}^{0(k)} \widehat{p}_{+j}^{0(k)} = x_{i+}^{(k)} x_{+j}^{(k)}/(kN)^2$. With the assumption (6), we could suppose that the Pearson Chi-square statistic $X^{2(k)}$ of the survey research approximates to the k times X^2 statistic value of the pilot study, That is,

$$X^{2(k)} = k X^2 + o_p(1). \tag{7}$$

We could say that the assumption (7) is reasonable because the main research will be surveyed in near future under the same situation of population of the pilot study.

Hence we might find that the p-value, $p(\theta^{(k)})$, which is the probability of that X^2 random variable is greater than $\theta^{(k)}$ is very much similar with $p(k\theta)$, because the χ^2 distributions of both the pilot study and the prospective survey have the same degree of freedom $(I-1)\times(J-1)$. Therefore, with the strong assumption (7) one can get

$$p(\theta^{(k)}) = p(k\theta) + o_p(1) . \tag{8}$$

By using the linear relationship between k and $\log[p(k\theta)]$ discussed in Section 2, we can expect an appropriate value of k corresponding to a significant level α. In other words, a sufficient sample size of a survey research could be estimated with an appropriate value of k.

[Example] The following data in <Table 3> is taken from the 1984 General Social Survey of the National Data Program in the United States as quoted by Norušis (1988) The variables are income and job satisfaction. Income has levels less than \$6,000 (denoted < 6), between \$6,000 and \$15,000 (6-15), between \$15,000 and \$25,000 (15-25), and over \$25,000 (>25). Job satisfaction has levels very dissatisfied (VD), and little dissatisfied (LD), moderately satisfied (MD), and very satisfied (VS). We treat VS as the high end of the job satisfaction scale. <Table 3> contains the estimated expected frequencies for H_0 : independence. Two goodness-of-fit statistics have values of X^2 =11.99 and G^2=12.03, based on *d.f.*=(4-1)(4-1)=9. Both statistics yield a p-value of 0.21 so that this data does not show strong evidence of association between income and job satisfaction. This result is not what the researchers want, so that let us assume that the work analyzed previously is regarded as a pilot study and the next prospective survey research is necessarily demanded in near future under invariant situation compared with the pilot study. Then in order to decide sample size of the prospective survey, let us use the linear relationship between k and $\log[p(k\theta)]$.

<Table 4> shows observed p-values, $p(k\theta)$, and predicted log p-values, $\log \widehat{p}(k\theta)$=

3.08 - 4.33 k, which are obtained from the least squared regression line in (4) when the values of k are greater than 1 and less than 3. When k =1.4 (kN= 1,260), one can find that the predicted p-value is little greater than a significant level α (=0.05, for example). An appropriate sample size of a future survey, kN, could be predicted. Therefore, if the sample sizes of a prospective survey is selected to be greater than 1,300 ($kN \geq 1,300$) with some appropriate assumptions, the researchers will collected data and could obtain some conclusions which they want.

<Table 3> Job satisfaction data

	VD	LD	MD	VS	total
< 6	20 (14.175)	24 (24.693)	80 (72.935)	82 (94.198)	206
6 - 15	22 (19.887)	38 (34.642)	104 (102.32)	125 (132.15)	289
15 - 25	13 (16.171)	28 (28.169)	81 (83.202)	113 (107.46)	235
> 25	7 (11.767)	18 (20.497)	54 (60.543)	92 (78.193)	171
total	62	108	319	412	901

<Table 4> observed p-values $p(k\theta)$ and predicted log p-values

k	$p(k\theta)$	$\hat{p}(k\theta)$	$\log p(k\theta)$	$\log \hat{p}(k\theta)$	residuals	kN
1	0.21395419	0.287086	-1.54199	-1.24797	-0.29402	901
1.1	0.15430696	0.186173	-1.86881	-1.68108	-0.18773	991
1.2	0.10923242	0.120731	-2.21428	-2.11419	-0.10009	1081
1.3	0.07606670	0.078293	-2.57614	-2.54730	-0.02885	1171
1.4	0.05220787	0.050772	-2.95252	-2.98040	0.02788	1261
1.5	0.03537290	0.032925	-3.34181	-3.41351	0.07170	1352
1.6	0.02369135	0.021352	-3.74265	-3.84662	0.10398	1442
...						
3.0	0.00004021	0.000050	-10.12150	-9.91014	-0.21136	2073

4. Application II : Estimation of least squared regression line

Once a goodness-of-fit test statistic value of a certain contingency table is obtained, the values of $\log p(k\theta)$ can be calculated by using the results in Section 3. And one might predict the regression line from the data $\{(k, \log p(k\theta))\}$. Then we compare $\{\log p(k\theta)\}$ with $\{\log \hat{p}(k\theta)\}$. We would like to take a look some values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$

at some cases of the various degrees of freedom and different values of corresponding goodness-of-fit test statistics.

First of all, given the interval of $k \in [1 \le k \le 3]$, 3×3, 3×4, and 4×4 contingency tables are considered. Each corresponding degrees of freedom are 4, 6, and 9, respectively.

4.1 $\{1 \le k \le 3\}$ is given

[1] Case 1 : $X^2 = 4.73$ $(=\theta)$, d.f.= 4 $(=n)$

Since the degree of freedom is an even number, one get values of $p(k\theta)$ using equation (2) such that

$$\begin{aligned}\log[p(k\theta)] &= -\frac{k\theta}{2} + \log\left[\sum_{i=0}^{\frac{n}{2}-1}\frac{(k\theta/2)^i}{i!}\right] \\ &= -\frac{4.73k}{2} + \log\left[\sum_{i=0}^{\frac{4}{2}-1}\frac{(4.73k/2)^i}{i!}\right] \\ &= -\frac{4.73k}{2} + \log\left[1+\frac{4.73k}{2}\right] .\end{aligned}$$

With the values of $\{(k,\ \log[p(k\theta)])\ ;\ 1 \le k \le 3\}$, the predicted least squared regression line is obtained as

$$\log[\hat{p}(k\theta)] = b_0 + b_1 k = 0.8535 - 1.9356 \cdot k .$$

Values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are found by using iterative estimation method such as 1.349 or 2.566.

[2] Case 2 : $X^2 = 4.7512$, d.f.= 6

$$\log[p(k\theta)] = -\frac{4.7512k}{2} + \log\left[1+\frac{4.7512k}{2}+\frac{(4.7512k/2)^2}{2}\right] ,$$

$$\log[\hat{p}(k\theta)] = 1.1173 - 1.5464 \cdot k .$$

The values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are 1.355 or 2.569, which are equivalent to those of Case 1.

[3] Case 3 : $X^2 = 11.99$, d.f.= 9 (based on <Table 3> of Job satisfaction data)

$$\log[p(k\theta)] = \log\left[1-\frac{2}{\sqrt{\pi}}\int_0^{\sqrt{11.99k}} e^{-x^2}dx + \sqrt{\frac{2}{\pi}}\ e^{-\frac{11.99k}{2}}\ \sum_{i=1}^{4}\frac{2^{i-1}\,(i-1)!\,(11.99k)^{\frac{2i-1}{2}}}{(2i-1)!}\right]$$

$\log[\hat{p}(k\theta)] = 3.0831 - 4.33108 \cdot k$

The values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are 1.349 or 2.563, which are almost equivalent to those of Case 1 and 2.

One finds that, for Case 1 to 3 of Section 4.1, rounded values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are almost identical even though values of goodness-of-fit statistic and degrees of freedom are different. Now we consider another interval of $k \in [1 \le k \le 5]$. The same 3×3, 3×4, and 4×4 contingency tables as those in Section 4.1 are reconsidered.

4.2 { $1 \le k \le 5$ } is given

[1] Case 1 : $X^2 = 5.127$, d.f.= 4

$$\log[p(k\theta)] = -\frac{5.127k}{2} + \log[1 + \frac{5.127k}{2}]$$

$$\log[\hat{p}(k\theta)] = 1.1308 - 2.2428 \cdot k$$

The values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are 1.684 or 4.073.

[2] Case 2 : $X^2 = 4.7512$, d.f.= 6

$$\log[p(k\theta)] = -\frac{4.7512k}{2} + \log[1 + \frac{4.7512k}{2} + \frac{(4.7512k/2)^2}{2}]$$

$$\log[\hat{p}(k\theta)] = 1.5245 - 1.7554 \cdot k$$

The values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are 1.697 or 4.079, which are almost equivalent to those of Case 1.

[3] Case 3 : $X^2 = 11.99$, d.f.= 9 (based on <Table 3> of Job satisfaction data)

$$\log[p(k\theta)] = \log[1 - \frac{2}{\sqrt{\pi}} \int_0^{\sqrt{11.99k}} e^{-x^2} dx + \sqrt{\frac{2}{\pi}}\, e^{-\frac{11.99k}{2}} \sum_{i=1}^{4} \frac{2^{i-1}\,(i-1)!\,(11.99\,k)^{\frac{2i-1}{2}}}{(2i-1)!}]$$

$$\log[\hat{p}(k\theta)] = 3.9767 - 4.7906 \cdot k$$

The values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are 1.679 or 4.067, which are almost equivalent to those of Case 1 and 2.

From Case 1 to 3 of Section 4.2, rounded values of k satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are almost equivalent even though values of goodness-of-fit statistic and degrees of freedom are different, which is the similar arguments of Section 4.1. From these above results, surprisingly, we found that, once the interval of k is determined, the values of k satisfying

$\log p(k\theta) = \log \hat{p}(k\theta)$ are equivalent approximately no matter what the degrees of freedom and value of corresponding goodness-of-fit test statistic are. Therefore, one can select two rounding-off values of k (say k_1, k_2), which are free of the values of goodness-of-fit statistic and degrees of freedom.

Moreover, the regression line might be estimated with the corresponding two points $\{(k_1, \log[p(k_1\theta)]), (k_2, \log[p(k_2\theta)])\}$ as the following :

$$\log[\hat{p}^*(k\theta)] = \frac{\log[p(k_1\theta)] - \log[p(k_2\theta)]}{k_1 - k_2}(k - k_1) + \log[p(k_1\theta)] . \quad (9)$$

For example, we already obtained the regression line (4) by the least squared method such that $\log[\hat{p}(k\theta)] = 3.08 - 4.33\,k$ for Job satisfaction data in <Table 3> with many values of $\{(k, \log[p(k\theta)])\}$. For a given interval $\{1 \le k \le 3\}$, one could select two points of k which are 1.35 and 2.57. Then alternative regression line might be estimated as the following :

$$\begin{aligned} \log[\hat{p}^*(k\theta)] &= \frac{\log[p(1.35\theta)] - \log[p(2.57\theta)]}{1.35 - 2.57}(k - 1.35) + \log[p(1.35\theta)] \\ &= \frac{-2.76 + 8.05}{1.35 - 2.57}(k - 1.35) - 2.76 \\ &= 3.07 - 4.33\,k . \end{aligned} \quad (10)$$

We might say that equation (10) is equivalent to equation (4), so that the regression line which is useful to determine sample size for survey research could be estimated by using the simple method of (9) with only two data points.

5. Conclusion and further study

When a random variable X^2 follows the Chi-square distribution with n degrees of freedom, the p-value of the Chi-square distribution is defined as, for given θ, $p(k\theta) = \Pr(\chi^2 > k\theta)$, $k \ge 1$. In this paper, our first finding is that the relationship between k and $\log[p(k\theta)]$ is almost linear. With this result, we could estimate a regression line by using the least squared method from data $\{(k, \log[p(k\theta)]) ; k \ge 1\}$.

Once the interval of k is determined, one obtains some data $\{(k, \log[p(k\theta)]) ; k \in$ some interval$\}$. We found that two values of k (say k_1, k_2) satisfying $\log p(k\theta) = \log \hat{p}(k\theta)$ are all

the same, no matter what the degrees of freedom and value of corresponding goodness-of-fit test statistic are. Hence alternative regression equation might be formulated in (9) with only two data points.

Note that the equation, $\log p(k\theta) = \log \hat{p}(k\theta)$, is the form of

$$\log[(n-1)\text{th polynomial equation of } k] = b_0 + b_1 k .$$

The values of k satisfying equation of the above form can not be obtained by general algebraic solution methods but are found by using iterative estimation methods.

As one of many applications of the theories, one might think an estimation method of sample size for survey research, which is discussed in Section 3. Under some regularity conditions and the linear relationships between k and $\log[p(k\theta)]$, one could predict an appropriate value of k corresponding to a given significant level. Then a sufficient sample size for a prospective survey research could be estimated.

For testing $H_0: \mu = \mu_0$, which is shown at any undergraduate textbooks, we might consider that sample sizes of a pilot study and a survey are N and kN, respectively. One knows that the following random variables

$$t_{N-1} = \frac{\overline{X}_N - \mu_0}{S_N/\sqrt{N}} \text{ and } t_{kN-1} = \frac{\overline{X}_{kN} - \mu_0}{S_{kN}/\sqrt{kN}}$$

follow t distributions with $N-1$ and $kN-1$ degrees of freedom, respectively. With similar arguments of Section 2 and 3, assume that the characteristic of the population does not change at either situations of a pilot study and a prospective survey. Then it is easy to get the results that both expectations of $\overline{X}_N$ and $\overline{X}_{kN}$ are μ and those variance estimates converge to 0 as each sample size increases. Now one could assume that as we did in <Theorem>,

$$\overline{X}_N - \overline{X}_{kN} \to 0 \quad \text{in prob} ,$$
$$\frac{S_N}{\sqrt{N}} - \frac{S_{kN}}{\sqrt{kN}} \to 0 \quad \text{in prob} .$$

One might say that value (c_N) of the above random variable t_{N-1} gets very much similar as that (c_{kN}) of t_{kN-1} under assuming invariance in their population. That is, $t_{N-1} = t_{kN-1} + o_p(1)$, which looks different from equation (7). For given values c_N, c_{kN}

of t test statistics, the corresponding p-values are defined as

$$p(c_N) = \Pr(t_{N-1} > c_N)$$

for a pilot study, and

$$p(c_{kN}) = \Pr(t_{kN-1} > c_{kN})$$

for a prospective survey. Since the degrees of freedom of t test statistics corresponding to p-values, $p(c_N)$ and $p(c_{kN})$, are different, we cannot argue that $p(c_N) = p(c_{kN}) + o_p(1)$. Nonetheless, when both sample sizes N and kN get larger, it is trivial to show that $p(c_N) = p(c_{kN}) + o_p(1)$ by some monte carlo studies, which is similar assumption as equation (8).

For further studies, this work could be extended to test $H_0 : \mu_1 = \mu_2$ in order to determine sample sizes of future researches. And for F test statistics in ANOVA tables, similar arguments could be explored.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*, John Wiely & Sons, New York.

[2] Bishop, Y. M. M., Fienberg, S. E. W., and Holland, P. W. (1975). *Discrete Multivariate Analysis : Theory and Practice*, The MIT Press.

[3] Christensen, R. (1990). *Log-Linear Models*, Springer-Verlag.

[4] Cressie, N. A. C., and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *Journal Royal Statistical Society*, B46, 440-464.

[5] Norušis, M. J. (1988). *SPSS^x Advanced statistics Guide* (2nd edition), McGraw-Hill, New York.

[6] Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables and calculation of P., *Journal of the Royal Statistical Society*, Series A, 85, 87-94.

[7] Johnson, N. L., and Kotz, S. (1970). *Continuous univariate distributions-1*, John Wiely & Sons, New York.

[8] Park, S. H., and Huh, M. Y. (1983). *Computational Statistics,* Kyung-Moon Sa, Seoul.

[9] Wilson, E. B., and Hilferty, M. M. (1931). The distribution of chi-square, *Proceedings of the National Academy of Sciences*, Washington, 17, 684-688.

[2002년 6월 접수, 2002년 10월 채택]