한국통계학회논문집 제2권 2호
1995년 12월 pp. 176-183

On Combining MOS and Histogram in a Subjective Evaluation Method[1)]

Sehyug Kwon[2)]

Abstract

Mean opinion score (MOS) method has been used in many areas to quantify opinions of respondents not only in survey research but in evaluating the parameters of population that are not measurable or are technically hard to be measured. Histogram is an important graphical technique because of the role it plays in describing categorical data as well as quantitative. In MOS method, subjective opinions of respondents are quantified by opinion scores and the arithmetic means of opinion scores have been used to describe the interesting population. Since opinion scores are polytomous, the values of arithmetic means have little meanings. In this paper, cumulative percentage curves as a function of the means of opinion scores are derived by combining means of opinion scores and histograms. It is proposed for better interpretation to opinion scores in MOS method, one of subjective evaluation methods.

1. Introduction

Survey research involves obtaining information directly from a group of individuals. Dane (1990) mentioned that there are three different types of information that may be obtained from survey; fact, opinion, and behavior. Fact is a phenomenon or characteristic available to anyone who knows how to observe it, but opinion is an expression of a respondent's preference, feeling, or behavioral intention. Behavior refers to an action completed by a respondent. There are probably as many different reasons for conducting surveys as there are surveys. Surveys are frequently conducted for the purpose of making descriptive assertion about some population: discovering the distribution of certain attributes (Babbie 1973). In survey research, opinions are used to describe the preference or favor of an interesting population. Then, they are obtained through questionnaire that would be answered by sampled or specified respondents from the interesting population. The variable corresponding to opinions that are expressed by respondents usually is categorical or qualitative. Sometimes opinions can be or should be quantified. When a survey is carried out to evaluate the parameter of population

1) This paper was supported by the research grant from Hannam University.
2) Department of Applied Statistics, Hannam University, Taejon, 300-791, KOREA

that is not measurable or technically hard to be measured, opinions of respondents should be quantified.

For example, quality of service (QOS) is one of the important concepts in telecommunications. In general, the quality has been defined as "fitness for use" or "loss from expectation". The latter definition is more preferable in the custom view (Clark 1991). The network performance that is included in the quality of service is measured by bit error ratio (BER) in data communications which is technically measurable. There are some technical difficulties in measuring BER, including the fact that BER does not directly affect the QOS in telecommunications. Moreover, QOS has a more direct impact on customers' evaluations of service quality. Therefore, QOS in telecommunications is frequently measured by inquiring and quantifying the subjective opinions of users through questionnaire.

When data set is categorical in nature as well as quantitative, the histogram is an important graphical technique because of the role it plays in statistical inference (Ott 1993), which can show the central tendency and spread of data set. In survey research, especially inquiring opinions of specified respondents through questionnaire, relative frequency histograms are useful graphs to describe and summarize the collected data set. The relative frequency histogram for an answered question is constructed by drawing two axes: a horizontal axis labeled with the specified question items and a vertical labeled with the relative frequencies of each question items. It shows the proportions of respondents who prefer a given question item.

When the question items can or should be quantified as mentioned, two important numerical measures, means (the measure of center) and standard deviation (the measure of variability), are common and useful statistical summaries for surveyed data set. Mean opinion score (MOS) method, one of subjective evaluation methods, has been used widely to quantify respondent's opinion not only in survey research but in evaluating the parameters of population that are not measurable or are technically hard to be measured. In MOS method, opinions of respondents to a given question are usually quantified by five scaled scores (opinion scores) from score 1 to score 5 and the arithmetic mean and standard deviation of opinion scores are computed. MOS method is simple and easy to quantify subjective opinions of respondents and summarize the surveyed data set with the numerical summaries. Therefore, it has been widely used in many area.

Since opinion scores are polytomous in MOS method, using the arithmetic mean for describing data has some problems that will be discussed in the next section. Using normal approximation to histogram, cumulative percentage curves are derived in this paper to combine MOS method and histogram. In cumulative percentage curves, means of opinion scores at the horizontal axis and histogram at the vertical axis are plotted in the same graph. The image quality of G3 facsimile in PSTN (Public Switched Telephone Network) is measured as an application of the proposed cumulative percentage curves.

2. MOS and its problems

When the survey is carried out and MOS method is used to quantify the opinions of respondents, each question in questionnaire has generally five categories (question items) to inquire opinions of respondents as follow: "unsatisfactory", "poor", "fair", "good", "excellent". Those categories are quantified by opinion scores. Rating opinion score 1 represents the unsatisfactory category, 2 represents poor, 3 represents fair, 4 represents good, and 5 represents excellent. Let Pi be the proportion of opinion scores for i=1, 2, 3, 4, 5 that are answered by specified respondents. The mean and standard deviation of opinion scores are respectively calculated by

$$Mean = \sum_{i=1}^{5} iP_i$$
$$Std = [\sum_{i=1}^{5} i^2 P_i - (Mean)^2]^{(1/2)} \quad (1)$$

The values of arithmetic means have been used to describe the preference or favor of the population. Suppose that the subjective evaluation of consumers to product A is surveyed. The question that is answered by consumers would be "What do you think the quality of product A is?". Then question items for the given question would be the followings: (1) unsatisfactory, (2) poor, (3) fair, (4) good , and (5) excellent. After survey, the sample mean of opinion scores of sampled consumer is computed by (1). In MOS method, it has been said that the quality of product A being evaluated by consumers is the mean value of opinion scores of respondents.

Suppose that the sample mean value of opinion scores to product A is computed 4.1. What can we say about 4.1 in MOS method? It can be said "Just better than good". How about 3.9? we just can say "Just worse than good" or "much better than fair" from the interpretation of MOS method. Since quantified opinions are polytomous, the value of mean of opinion scores has little meanings. There are some problems in comparing two populations by MOS method. Suppose that the sample mean of the opinion scores to product A is computed 4.1 and that of opinion scores to product B is 3.9. When two population are compared, it can not be said that the quality of product A is higher on the average than that of product B unless the same group of respondents is surveyed. Moreover, the difference between two sample mean values in even the same population has no meaning except relative magnitude.

3. Derivation of cumulative percentage curves

Opinions that are quantified by MOS can be summarized by histogram and numerical summaries, mean and standard deviation. If histogram and numerical summaries could be combined, it would give us better summary on the interesting population and clear out

drawback of both methods somewhat. For this purpose, the cumulative percentage curves is derived for plotting MOS and relative frequencies in the same graph.

Denote g the combination of factors that may affect subjective opinions of respondents for g=1, 2, ..., G. The statistical significance tests of factors in models for polytomous was mentioned in McCullagh and Nelder (1989). Even when the factors are not significant at all, the cumulative percentage curves is still useful without the loss, which will be seen in the next section. The score histogram for each group, g, can be approximated by a normal density curve with mean ($\overline{X}_g$) and standard deviation (S_g). The range of opinion scores is divided as follow: from minus infinity to 1.5 as 1 score, 1.5 to 2.5 as 2 score, 2.5 to 3.5 as 3 score, 3.5 to 4.5 as 4 score, and 4.5 to infinity as 5 score, and the mean and the standard deviation of normal density curve of the g-th group are computed by solving the following simultaneous equations:

$$Mean_g = 5 - \sum_{i=1}^{4} \Phi\left(\frac{i+0.5-\overline{X}_g}{S_g}\right)$$

$$Std_g = \left[25 - \sum_{i=1}^{4}(2i+1)\Phi\left(\frac{i+0.5-\overline{X}_g}{S_g}\right) - (mean_g)^2\right]^{(1/2)} \qquad (2)$$

$$where\ \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

The cumulative percentage curves as a function of the means of opinion scores can be derived to combine opinion scores and histogram by the following procedures:

(Step ①) Calculate the sample mean and standard deviation of each group:

Sample means and standard deviations are calculated by using the equations, (1). The mean ($\overline{X}_g$) and standard deviation (S_g) of a normal density curve approximating the score histogram of each group are computed by solving the simultaneous equation system, (2).

(Step ②) Estimate an overall standard deviation, S , and adjust by using S:

Let n_g be the number of observations in the g-th group. Then, the overall standard deviation can be estimated by pooling S_g^2's as follow:

$$S = \left[\sum_{g=1}^{G} \frac{(n_g-1)S_g^2}{N}\right]^{(1/2)}, where\ N = \sum_{g=1}^{G}(n_g-1) \qquad (3)$$

After replacing S in the place of S_g in the first equation of (2), $\overline{X}_g$ is computed again.

(Step ③) Fit the cumulative percentage curves:

Using the adjusted means in step (, cumulative percentage curves subjects at a mean of opinion scores are fitted to polynomial curves of appropriate degree by a least squares fit technique. A percentage curve for the question item, "excellent", is fitted as follow: The adjusted means, the horizontal axis, and the relative frequencies of "excellent" at each histogram where the adjusted means are respectively derived, the vertical axis, are plotted. a percentage curve is fitted on a least squares error basis. The percentage curves can be fitted using polynomial curves of degree. A percentage curve for "more than and equal to good" is fitted with the cumulative relative frequencies of "excellent" or "good". With the same procedure, the cumulative percentage curves are fitted.

(Step ④) Interpret the fitted cumulative percentage curves:

The population is described by the point where the vertical line at the grand mean of opinion scores and a cumulative percentage curve meet. The use of fitted cumulative percentage curves for interpretation will be shown in the next section.

4. An application

In facsimile transmission, one form of data communication in PSTN, it seems likely that frame error ratio (FER) would be considered instead of BER to measure the image quality of G3 facsimile transmission. It has some technical difficulties to measure including the fact that FER does not directly affect the image quality as mentioned by TSS (1994). Therefore, the image quality of G3 facsimile is measured by surveying subjective evaluation of facsimile users using MOS method.

An ITU-TS test chart No. 2 of 297 mm length and 210 mm width contains elements permitting quantitative evaluation of distortion and character groups intended for evaluation of the readability of the transmitted documents. The chart is recommended in ITU-TS (1988) to measure the image quality of G3 facsimile. About 10 questions are normally asked to calculate the overall quality of transmitted document. However, Kwon and Hwang (1994) showed that the following question could replace questions about all of the individual elements to measure the image quality of transmitted documents: "What do you think the overall quality of this transmitted chart is?". Therefore, the opinions of facsimile users to the single question are quantified by opinion scores for measuring the quality of G3 facsimile.

Even though the relationship between FER and the image quality of G3 facsimile does not seem linear, the image quality deteriorates as FER increases. Therefore, the following can be considered main factors that may affect the image quality of facsimile: the number of links, line traffic, and condition of transmission lines. In general, as the number of links involved in data transmission increases, the FER goes up. There are less than or equal to 4 links in

Korea unless detour routes are used. When lines are busy, it seems very likely that FER goes up. Therefore, the line traffic should be considered a factor that affects the image quality. A test time would be split into a busy period and a non-busy period. From 9 a.m. to 12 p.m. would be considered busy hours (BH). The other hours would be non-busy hours (NB). There are many variables that affect the condition of lines during transmission, for example, the type or characteristic of trunks, circumstances affecting transmission trunks, and so on. It is hard to consider the condition of lines a factor. Therefore, the two main factors, the link effect (L) and the traffic effect (T), and their interaction effect are considered factors that may affect the image quality of G3 facsimile. The combinations (groups) of two factors used are the following:

$$\begin{cases} L = 1, 2, 3, 4 \\ T = BH, NB \end{cases}$$

Sample data have been collected to measure the quality of facsimile in real transmission lines as follows: According to the groups, some major cities in Korea are selected for document transmission and ITU-TS test chart No. 2s are transmitted between selected cities. Sixty-five facsimile users are selected to respond questionnaire. Three test charts are transmitted for each combination and evaluated by sixty-five facsimile users answering the question, "What do you think the overall quality of this transmitted test chart is?". Each question item in the questionnaire has five categories scaled, for example, from 5 to 1 corresponding to excellent, good, fair, poor and bad.

The significance of the main effects and interaction effect are tested as suggested in McCullagh and Nelder (1989). They are all insignificant at α=0.05. Computations for deriving the cumulative percentage curves are summarized in Table 4.1 and the estimated overall standard deviation by the equation, (3), are 0.7692. The percentage curves of cumulative subjects at a mean of opinion scores are derived as seen in Figure 4.1. Since the effects are not significant, the range of the adjusted means from 2.9 to 3.4 is not wide as expected. The percentage curves are fitted to a linear function of the adjusted mean of a normal density. The lines mark the fitted percentage cumulative curve. For example, the cumulative percentage of the image quality being evaluated "good or excellent" at an mean of opinion scores value can be obtained from the third line with dot symbols.

Table 4.1 Summary of computation

Steps	Statistics \ Groups	1	2	3	4	5	6	7	8
①	$\overline{X}_g$	3.332	3.117	3.232	3.221	2.929	2.939	3.068	3.395
	S_g	.7712	.7949	.7300	.6736	.8172	.7084	.8106	.8326
②	$\overline{X}_g$	3.334	3.121	3.232	3.222	2.929	2.940	3.071	3.394

The fitted cumulative percentage curves, Figure 4.1, are interpreted as follows in evaluating the quality of facsimile: The grand mean opinion score of sampled facsimile users is 3.154. Therefore, 30% of facsimile users consider the quality of the facsimile as "good or excellent". And 80% of them are evaluating the quality as "fair, good or excellent". Therefore, it can be concluded that the image quality of G3 facsimile is evaluated as at least as "fair" in the present PSTN. If the proportion of facsimile users that evaluate the image quality as "good or excellent" is targeted 40%, then the observed mean of opinion scores of specified or sampled respondents from another survey should be greater than and equal to 3.295. This interpretation of cumulative percentage curves can be applied in comparing two populations.

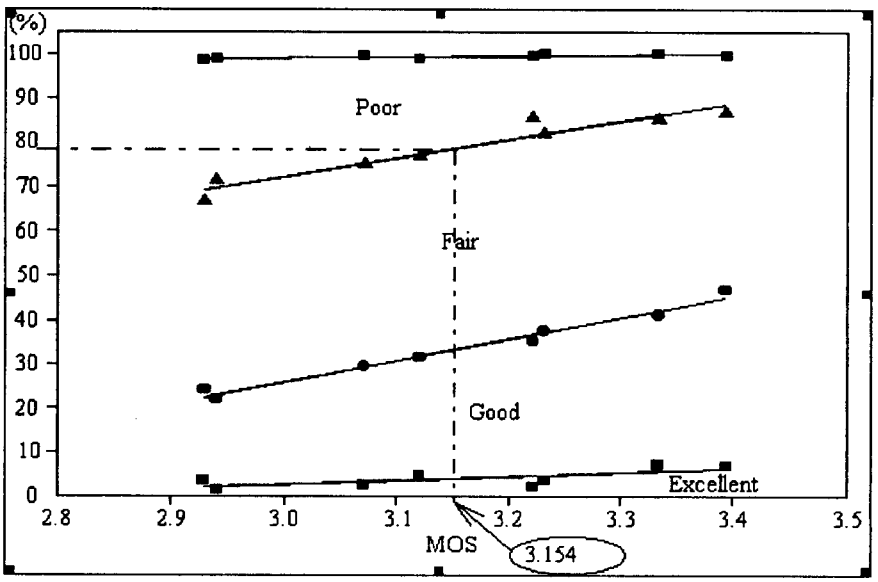

Figure 4.1 The cumulative percentage curves subjects at an mean of opinion scores

5. Conclusion

Histogram is an useful graphical technique to show frequency distribution of the population whose variable is categorical as well as quantitative and MOS is an widely used method to quantify subjective opinion of respondents among subjective evaluation methods. Therefore, the cumulative percentage curves are derived in this paper by combing MOS method and histogram. They give better interpretation to opinion score for describing the nature of the population in the case of one population and are also effective in comparing two populations.

References

[1] Babbie, E. R. (1973). Survey Research Methods, Wadsworth publishing company.

[2] Clark, R. (1991). Network Performance; The Custom View, *ITU Telecommunications Journal*, Vol. 58, 360-363.

[3] Dane, F. C. (1990). Research Methods, Thompson Information/Publishing Group.

[4] Guilford, J. (1954). Psychometric Method, 2nd Edition, McGraw-Hill.

[5] ITU-TS (1984). Handbook on Quality of Service, Network Management, and Network Maintenance, ITU, Geneva.

[6] ITU-TS (1988). Terminal Equipment and Protocols for Telemetric Services, Series T.21, ITU, Geneva, 71-75.

[7] Kwon, S. H. and Hwang, G. (1994). A study on testing the image quality of facsimile, Electronics and Telecommunications Trends, Vol. 8, Number 4.

[8] McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, Chapman and Hall, London.

[9] Ott, R. L. (1993). An introduction to statistical methods and data analysis, 4th Edition, Wadsworth publish company.

[10] Telecommunication Standardization Sector (1994). Quality as corrupted by transmission -induced scan line errors, ITU Draft Recommendation E.456.