DIFFERENT APPROACHES USING THE NORMAL AND THE EXPONENTIAL DISTRIBUTION IN THE EVALUATION OF THE CUSTOMER SATISFACTION

Antonio LUCADAMO[1]
PhD, Assistant Researcher, Tedass,
University of Sannio, Benevento, Italy

E-mail: alucadam@unina.it

Giovanni PORTOSO[2]
PhD, Associated Professor, SEMEQ Department, Faculty of Economics,
University of Eastern Piedmont "A. Avogadro", Novara, Italy

E-mail: portoso@eco.unipmn.it

Abstract: *The Customer Satisfaction is generally evaluated using the data collected with questionnaires. The data are organized on an ordinal scale and, for this reason, it's convenient to transform them in pseudo-interval support. The psychometric methods used for this transformation generally hypothesize that the latent variable has a normal distribution. Sometimes, particularly when the frequencies are concentrated on the left extreme or on the right extreme of the distribution, this assumption brings to preposterous results. In these cases the use of other types of distribution, as, for example the exponential distribution, is preferable. In this paper we show how the results of a survey can change using the normal distribution, the exponential distribution or the two distributions alternatively. We use, in fact, the results coming from the different transformations, to apply a multilevel model.*

Key words: *customer satisfaction; normal distribution; exponential distribution; multilevel models*

1. Introduction

One of the problem of the Customer Satisfaction is the quantification that converts on a metric scale the judgements about services or products. A simple technique is the so-called "direct quantification": this technique hypothesizes that the modalities of a qualitative character are at the same distance, but this hypothesis is not respected in many situations (Marbach, 1974). For this reason it is preferable to use an alternative technique, the "indirect quantification", that consists in assigning real numbers to the categories of the qualitative variable. In this type of quantification the numbers are not equidistant but they depend on a latent variable. Different measurement techniques have been developed during the years (Thurstone, 1925, Guilford, 1936, Torgenson, 1958) based on the hypothesis that the model is normally distributed. This assumption can be realistic in a psychometric field, but it is not always valid in the Customer Satisfaction, especially if the judgements are all extremely positive or extremely negative. More recent techniques have been proposed, based for example on the use of logit and probit models, on structural equation models and so on. In next section we introduce the psychometric quantification, underlining the problems that can arise in some situations; then we show how the use of another kind of quantification can solve these pitfalls and, in the following paragraphs we propose the use of a combined technique, showing the results that we obtain on real data.

2. The psychometric quantification

In the psychometric quantification, the modalities x_i ($i = 1,2,\dots,r$) of a qualitative variable X, are associated to the values of a quantitative latent variable Z, normally distributed. Let $F(i)$ be the cumulative relative frequency, corresponding to x_i and let $\Phi^{-1}[F(i)]$ the inverse of the cumulative distribution function, the quantile z_i associated to x_i can be expressed as $z_i = \Phi^{-1}[F(i)]$. To obtain the new scores, we simply calculate the expected values $E(Z_i)$ over all the X variables in the data-set. The assumption of the normal distribution, when the frequencies are prevalently on the left extreme or on the right extreme of the distribution, leads to strange results. In fact the scores will be negative if the modalities are almost on positive side and vice-versa (the results in Table 1 can help to understand the situation) (Portoso, 2003a).

Table 1. Quantification with the normal distribution of the judgements given on two different services

Judgements	Frequencies of the first service	Frequencies of the second service
Very negative	350	10
Negative	80	20
Indifferent	40	40
Positive	20	80
Very positive	10	350
Totals	500	500
Expected quantile	0.0729	-0.0729

It is easy to see that the first service has many negative judgements, so the frequencies are prevalently on the left side of the distribution, but the expected quantile has a positive value. For the second service there is instead the inverse situation, in fact the frequencies are on the right side, but expected value of the quantile is negative.

This incongruity leads to use a distribution that could better express, in a numerical way, the categorical variables characterized by this particular structure. The exponential distribution seems to be the right solution.

3. The exponential quantification

In this section we show how to determine a quantification based on the negative and on the positive exponential distribution. Before introducing the new procedure, it's necessary to describe briefly the two cited distributions.

3.1. The negative exponential distribution

Let consider

$$\begin{cases}\psi(z)=\exp(-z) \quad if \quad (0\le z\le\infty)\\ \psi(z)=0 \qquad otherwise\end{cases} \tag{1}$$

where Z is a quantitative variables.

It can be assumed as the relative density function, in fact:

$$\int_0^\infty \psi(z)dz=\int_0^\infty \exp(-z)dz=1 \tag{2}$$

The mean and the variance are defined as follow:

$$E(Z)=\int_0^\infty z\psi(z)dz=\int_0^\infty z\exp(-z)dz=[-z\exp(-z)]_0^\infty-\int_0^\infty -\exp(-z)dz=1 \tag{3}$$

$$Var(Z)=\int_0^\infty z^2\exp(-z)dz-[E(Z)]^2=2-1^2=1 \tag{4}$$

The variable can be then standardized in the following way:

$$S=(Z-1)=Z-1 \tag{5}$$

with

$$\begin{cases}f(s)=\exp(-s-1) \quad if\,(-1\le s\le\infty)\\ f(s)=0 \qquad otherwise\end{cases} \tag{6}$$

This is a relative frequency density function, in fact:

$$\int_{-1}^{+\infty}\exp(-s-1)=1 \tag{7}$$

The cumulative distribution function is:

$$\begin{cases}\Psi(s)=\int_{-1}^{s}\exp(-t-1)dt=1-\exp(-s-1) \quad if \quad (-1\le s\le\infty)\\ \Psi(s)=0 \quad otherwise\end{cases} \tag{8}$$

3.2. The positive exponential distribution

Let consider

$$\begin{cases}\psi(y)=\exp(y) \quad if \quad (-\infty\le y\le 0)\\ \psi(y)=0 \qquad otherwise\end{cases} \tag{9}$$

that can be assumed as the relative density function, in fact:

$$\int_{-\infty}^{0} \psi(y)dy = \int_{-\infty}^{0} \exp(y)dy = 1 \qquad (10)$$

The mean and the variance are defined as follow:

$$E(Y) = \int_{-\infty}^{0} y\psi(y)dy = \int_{-\infty}^{0} y\exp(y)dy = [y\exp(y)]_{-\infty}^{0} - \int_{-\infty}^{0} \exp(y)dy = -1 \qquad (11)$$

$$Var(Y) = \int_{-\infty}^{0} y^2 \exp(y)dy - [E(Y)]^2 = 2 - (-1)^2 = 1 \qquad (12)$$

The variable can be then standardized in the following way:

$$P = (Y - (-1)) = Y + 1 \qquad (13)$$

The cumulative distribution function of P is :

$$\begin{cases} \Psi(p) = \int_{-\infty}^{p} \exp(t-1)dt = \exp(p-1) \quad if \quad (-\infty \le p \le 1) \\ \Psi(p) = 1 \quad otherwise \end{cases} \qquad (14)$$

3.3. The quantification

To build the scores, both for the negative exponential distribution and for the positive one, it's necessary to consider the relative frequencies $f(i)$ and the cumulative relative ones $F(i)$.

In this way we can define the following quantity (empirical distribution of cumulative frequencies):

$$G(i) = F(i-1) + f(i)/2 \quad i = 1, 2, \ldots, r \qquad (15)$$

If we consider the negative exponential distribution, we can compare formula (6) and (15) and we obtain the standardized quantile:

$$s_i = -1 - \ln[1 - G(i)] \qquad (16)$$

The same procedure can be applied for the positive distribution, using formula (13) and (15); in this case we will obtain the standardized quantile in the following way:

$$p_i = 1 + \ln[G(i)] \qquad (17)$$

To verify the importance, in particular situations, of using the exponential distribution, instead than the normal distribution, we can consider the value in table 2 in which there are the absolute frequencies about the judgments given to 8 different services by 500 judges.

Table 2. Absolute frequencies of the judgements given to 8 different services

	Services							
Judgments	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Very negative	496	470	20	180	20	10	2	1
Negative	1	16	50	50	120	20	4	0
Indifferent	1	8	360	40	220	40	6	0
Positive	1	4	60	50	120	350	96	0
Very positive	1	2	10	180	20	80	392	499
Total	500	500	500	500	500	500	500	500

In this table we can note that the services A and B received many negative judgments, the services F, G and H many positive judgments and the services C, D and E had a quasi-symmetric distribution.

This table is important to understand what happens when we apply the different kinds of quantification.

The results are shown in table 3.

Table 3. Quantiles associated to the relative cumulate frequencies centred on every judgment category in the hypothesis of exponential and normal distribution

Services		JUDGMENTS					Expected quantile
		Very negative	Negative	Indifferent	Positive	Very positive	
A	Exp neg.	-0.315	3.962	4.298	4.809	5.908	-0.274
	Norm	-0.010	2.457	2.576	2.748	3.090	0.012
B	Exp neg.	-0.385	2.124	2.912	3.828	5.215	-0.177
	Norm	-0.075	1.705	2.054	2.409	2.878	0.047
C	Exp neg.	-2.912	-1.408	0.307	0.917	0.990	0.093
	Norm	-2.054	-1.341	0	1.405	2.326	-0.001
D	Exp neg.	-0.802	-0.472	-0.307	-0.108	0.714	-0.114
	Norm	-0.915	-0.228	0	0.228	0.915	0
E	Exp neg.	-0.980	-0.826	-0.307	0.833	2.912	-0.056
	Norm	-2.054	-0.994	0	0.994	2.054	0
F	Exp pos.	-3.605	-2.219	-1.303	0.287	0.917	0.082
	Norm	-2.326	-1.751	-1.282	-0.025	1.405	-0.012
G	Exp pos.	-5.215	-3.828	-3.017	-1.120	0.502	0.091
	Norm	-2.878	-2.409	-2.097	-1.175	0.274	-0.067
H	Exp pos.	-5.908	-5.215	-5.215	-5.215	0.309	0.296
	Norm	-3.090	-2.878	-2.878	-2.878	0.003	-0.004
						General mean	0.002

Here we can see that, in a Customer Satisfaction analysis, when the frequencies are very high for judgments extremely positive or extremely negative, the use of the normal distribution is not an appropriate way to effectuate the quantification. Using the exponential distribution leads to better results, in fact we can see that for the services A and B that presented value extremely negative, we have that the expected value of the quantile is negative if we use the negative exponential distribution, while using the normal quantification it will be positive. For the services C, D and E there are no substantial differences between the use of the normal or of the exponential distribution, but the first one seems to be preferable; for this services in fact we had a symmetric distribution. For services G and H instead, the calculation of the expected quantile shows that the use of positive exponential distribution leads to positive values, while using the normal distribution we will have negative values.

Of course the observation of the expected quantile can not be the only instrument to decide if considering the normal distribution or the exponential distribution as latent variable, but we need an indicator that could help in the choice. A possible solution is given in Portoso (2003b) that introduces an useful index to decide which kind of distribution is better to apply in the different situations.

3.4. The EN index

The EN index is an indicator that assumes values between -1 and +1. The value -1 is assumed when all the frequencies are associated to the first modality (in this case we have maximum negative concentration), while when there is maximum positive concentration the value assumed by the index will be +1. If the frequencies are balanced in a symmetric way then the EN index will be equal to 0. The index has the following formulation:

$$EN = \sum_{i=1}^{r/2} (f_{r-i+1} - f_i)(r - 2i + 1)/(r-1) \tag{18}$$

where *r* is the number of modalities and if they are odd the value r/2 is round off to the smaller integer while f_i are, as already stated, the relative frequencies associated to the modality *i*, f_{r-i+1} are the frequencies associated to the opposite modality and $r - 2i + 1$ is the difference between the position of the two opposite modalities.

An alternative formulation of the index can be the following:

$$EN = 1 - 2\sum_{i=1}^{r-1} F(i)/(r-1) \tag{19}$$

that presents some similarities with the Gini index and where $F(i)$ have already been defined as cumulative relative frequency and r is the number of modalities of the qualitative variables. If the value of the index *EN* is close to 0, the use of normal distribution doesn't generate any problems, but if the absolute value of this index grows then the use of exponential distribution can lead to better results. The problem is to define a threshold to decide which distribution is better to apply. Portoso, with empirical attempts, showed that a value of the EN bigger than 0.2 in absolute value, indicates that the use of the exponential distribution is preferable to the normal one. In the following sections we first introduce briefly the multilevel models and then we verify what happens to the results of an analysis using the different kinds of quantification.

4. The multilevel models

Multilevel models suppose that in a hierarchical structure, the upper levels can influence the lower ones (Snijders, Bosker 1999). The basic model is the so called empty model defined as follows:

$$Y_{ij} = \gamma_{00} + U_{0j} + R_{ij} \tag{20}$$

In this formula there is a dependent variable Y_{ij} given by the sum of a general mean (γ_{00}), a random group effect (U_{0j}) and a random individual effect (R_{ij}). In this way the variability is divided in two parts: in fact, in this model it's assumed that the random variables U_{0j} and R_{ij} are mutually independent, normally distributed with zero mean and variances equal to τ^2 and σ^2. The total variance is then the sum of the two variances and we can compute the intra-class correlation coefficient:

$$\rho = \tau^2 / (\tau^2 + \sigma^2) \tag{21}$$

If this coefficient is significant, it is possible to effectuate a Multilevel Analysis (Hox, 2002). A first model is the Random Intercept Model that can be defined as follows:

$$Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + R_{ij} \quad \text{with} \quad \beta_{0j} = \gamma_{00} + U_{0j} \tag{22}$$

In the equation (20) if we consider the j subscript for the coefficient β_1 we will have the Random Slopes Model. In this case too we can see that there is a fix effect (γ_{10}) and a random ones (U_{1j}).

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij} \text{ with } \beta_{0j} = \gamma_{00} + U_{0j} \text{ and } \beta_{1j} = \gamma_{10} + U_{1j} \tag{23}$$

5. A case study

The application concerns a survey about Patient Satisfaction. The patients answered to 30 items relative to the services received during the staying in the hospital. They gave a score between 1 and 7 and furthermore they furnished information about the gender, the age and the education. To apply a multilevel model we need a variable relative to the second level and this is the experience of the head physician of the different wards. The aim is to verify if the Customer Satisfaction (CS) depends on the different variables and, above all, if the different quantifications leads to dissimilar results. For this reason we adopt the Normal Quantification, the Exponential Quantification and a Mixed Quantification (Normal or Exponential). The first two quantifications have already been illustrated, instead the third one is based on the use of the *EN* index. We use in fact the normal quantification for the items that have the *EN* index lower than a fixed threshold and the exponential distribution for the items with *EN* larger than the threshold. Furthermore, to compute the new scores, we used a geometric mean for the exponential quantifications because of lower sensitivity to extreme values and an arithmetic mean for the normal distribution. In Table 4 we show the number of the items transformed using the two distributions, according to the different thresholds, arbitrarily assumed and considering that it is possible a larger series of values.

Table 4. Number of the two distributions used according to the different thresholds

Threshold	0	0.2	0.3	0.4	0.5	0.6	0.7	0.8	1
N° of exponential	30	29	26	25	22	17	9	1	0
N° of normal	0	1	4	5	8	13	21	29	30

For all the criterions we then compute the overall CS as the sum of the new value that every individual has for the 30 items. In the building of the model, the only significant variable for the individual level is the age and the model that we adopt is a Random Intercept Model, so we can write:

$$CS_{ij} = \beta_{0j} + \beta_1 Age_{ij} + R_{ij} \quad \text{with} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}(Exp_j) + U_{0j} \qquad (24)$$

The results that we obtain are reported in Figure 1.

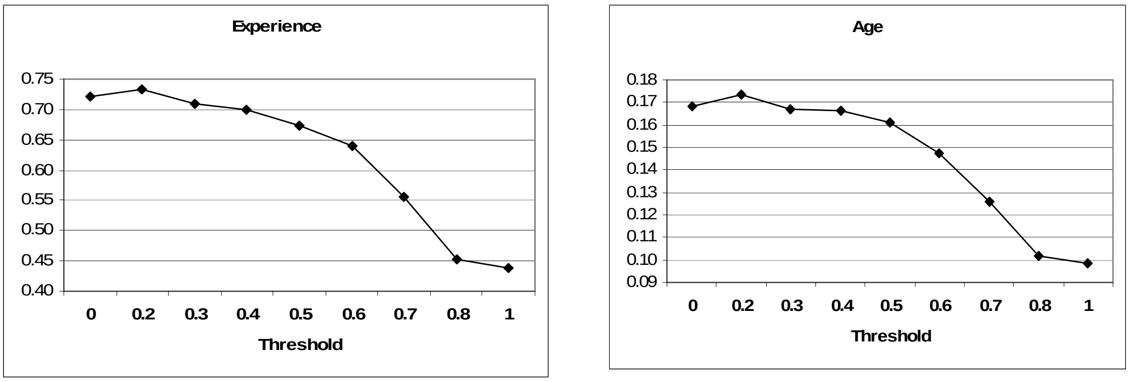

Figure 1. Coefficients of the two explicative variables with the different thresholds

We can note that the coefficients relative to the experience of the head physician (γ_{01}) and to the age of the patients (β_1) are both positive, so the CS is higher for older patients and for people that were nursed in department with expert doctors. Furthermore they increase if we consider an EN threshold that goes from 0 to 0.2 and then they both decrease considerably when the threshold is higher than 0.2, reaching a minimum by using only the normal as latent variable. Moreover they are all statistically significant and there are no substantial differences in the values of the t-ratio. The value of the *EN* = 0.2 is the value that Portoso (2003b) indicated as critical for the choice between exponential and normal distribution.

6. Considerations and perspectives

In the Customer Satisfaction or in the evaluation of other services there is a quantification problem that can not be solved using the direct quantification, because it doesn't answer to the reality. The use of the indirect quantification, with the assumption of a continuous latent distribution, is in this case preferable, but the choice can not be always the normal distribution. Using its standardization, the exponential distribution, negative or positive, has been assumed as an alternative to the normal when this one is not appropriate. The exponential distribution assures results that are more consistent with the shape of the empirical distribution and, furthermore, it guarantees distances between the modalities more adherent to the psychological continuum with which the judgments are expressed. The problem about the choice of the right distribution was discussed in an empirical way in a previous work; in this paper, the results that we obtain, introducing also a second step of the analysis, confirm the idea of using the exponential distribution instead than the normal one when the EN index is higher than 0.2 or smaller than -0.2. Obviously these are only results that comes from a restricted number of analyses and the definition of the threshold for the choice between normal and exponential distribution must be studied deeply. Furthermore some other indexes could be proposed and not only the use of normal and exponential distribution must be taken into account; our proposal is in fact to consider, in next works, the use of other distributions too.

References

1. Guilford, J.P. **Psychometrics Methods,** McGraw-Hill, New York, 1936, pp. 255-258
2. Hox, J.J. **Multilevel Analysis. Techniques and Applications,** Lawrence Erlbaum Associates, 2002
3. Marbach, G. **Sulla presunta equidistanza degli intervalli nelle scale di valutazione,** Metron, Vol. XXXII: 1-4, 1974
4. Portoso, G. **La quantificazione determinata indiretta nella customer satisfaction: un approccio basato sull'uso alternativo della normale e dell'esponenziale,** Quaderni di dipartimento SEMeQ, 53, 2003a
5. Portoso, G. **Un indicatore di addensamento codale di frequenze per variabili categoriche ordinali basate su giudizi,** Quaderni di dipartimento SEMeQ, 66, 2003b
6. Snijders, T. A. B. and Bosker R. J. **Multilevel Analysis. An introduction to basic and advanced multilevel modelling,** SAGE Publications, 1999
7. Thurstone, L. L. **A method of scaling psycological and educational tests,** J. Educ. Psychol., 16, 1925, pp. 443-451
8. Torgenson, W.S. **Theory and Methods of Scaling,** Wiley, New York, 1958

[1] **Antonio Lucadamo**, Philsophy Doctor in Statistics with the thesis "Multidimensional Analysis for the definition of the choice set in Discrete Choice Models"- department of Mathematics and Statistics, University of Neaples "Federico II". From 2007-2009 he had a post-doc research fellowship in Statistics about the Customer Satisfaction – department SEMEQ, University of Eastern Piedmont, Novara. Currently he has a post-doc research fellowship about the use of multivariate statistical methods in the spectrometry –Tedass, University of Sannio, Benevento. His research field are the Discrete Choice Models (Logit and Probit), the Multidimensional Analysis of Data (Principal Component Analysis, Cluster Analysis), the Multivariate Analysis for the definition of the sample size in clinical trials, the Rasch Analysis, the Quantification Methods and the Multilevel Regression for the evaluation of Customer Satisfaction.

[2] Graduated in Economics from the University of Bari in 1965 by a vote of 110 cum laude. Assistant Professor since 1971. In 1982 he became Associate Professor at the Faculty of Economics of Bari. He then taught at the Faculty of Economics in Turin. In the nineties, with 4 other colleagues, he formed the Faculty of Economics of Novara, where he currently teaches. Declared fit to full professor, he is not called by the Board of Faculty of Economics of Novara for lack of quorum. He has lectured in Anthropometry, Statistics, Statistical Methodology, Statistical Quality Control, Multivariate Statistical Analysis, Statistical Sampling, Theory of Probability and Mathematical Statistics, Market Research. His publications cover the methodological aspects and applications developed in the field of demographics and the financial sector, anthropometry, statistical analysis of the results of gambling, use of latent variables in the customer satisfaction, normalization of indices of association.