Statistical Analysis of Ordinal User Opinion Scores

Mu Mu, Andreas Mauthe
School of Computing and Communications
InfoLab21, Lancaster University
Lancaster, LA1 4WA, UK
Email: {m.mu, a.mauthe}@lancaster.ac.uk

Gareth Tyson
Department of Informatics
King's College London
London, WC2R 2LS, UK
Email: gareth.tyson@kcl.ac.uk

Eduardo Cerqueira
Federal University of Para
Belem, PA, Brazil
Email: cerqueira@ufpa.br

Abstract—**Data-sets derived from subjective experiments are often exploited to construct objective quality models using parametric statistics such as MOS and multiple regression. In this paper, using data type and normality tests, we verify that non-normally distributed user opinion scores in nominal or ordinal responses should *not* be analysed using parametric statistics. The paper introduces a number of non-parametric statistics for valid model building and parameter estimation based on user opinion scores. A set of modelling results are also presented to demonstrate the effectiveness of non-parametric statistics.**

I. Introduction

Due to the complexity of video services and the human visual system, subjective experiments are commonly conducted to investigate the quality of video content as perceived by human users. Often, these experimental results are exploited to construct objective quality models using statistical approaches and models. The modelling of qualitative user experiences using quantitative predictors (metrics) is a statistical inference process. Model specification, estimation of model parameters and verification of precision are the three aspects of a valid inference. Model specification, or formulation, in its widest sense, is conceptually more difficult than estimating model parameters [1]. Currently, the analysis of user opinion scores is commonly conducted using parametric statistics. This includes averaging user scores on each test condition as an absolute mean opinion score (MOS), building an objective model using general linear models and verifying model performance using Pearson correlation tests. These parametric statistics are then often exploited without the correct statistical validation of the model conditions. Therefore, using statistical models that do not match the characteristics of the data set usually leads to invalid analytical studies and objective models.

This paper introduces a group of methods and tools to examine the nature of experimental results against conditions of statistics. Data type tests, normality tests and independency tests are all applied on results of a typical subjective experiment conducted using standard test procedures. Test results verify that user opinion scores in the ITU recommended 5-point scale [2] are ordinal responses. Further, we find that the distribution of user scores is very unlikely to be normally distributed. Therefore, applying parametric statistics such as mean opinion scores and multiple regression on experimental results is not statistically correct. A number of non-parametric statistics are then introduced for valid statistical analysis where conditions for parametric statistics are not met. Cumulative logit model and maximum likelihood methods are recommended for model specification and parameter estimation. The paper also introduces corresponding methods to verify the goodness of a model fit and to calculate the confidence intervals of model predictions. A group of modelling results are also presented to demonstrate the effectiveness of modelling the distribution (rather than the arithmetic mean) of user opinion scores.

II. Background and Problem Space

This section reviews the current trend of subjective experiment and data analysis in the field of QoE research and defines the problem space as invalid utilisation of statistics.

A. Subjective Experiment and Data Analysis

Subjective user tests are vital in studying the qualitative user experience of a video service. Several international recommendations (e.g., [2], [3]) provide guidelines for valid subjective experiments. The guidelines specify multiple aspects of a test including the choice of rating scales, test environment, test materials as well as the type of communication with participants. The Video Quality Experts Group (VQEG) is a specialised group which has designed and conducted multiple dedicated test plans, such as the HDTV test plan [4] to benchmark the performance of objective models.

The modelling of qualitative user experience utilising quantitative impact metrics is a statistic inference process. Model specification, estimation of model parameters and estimation of precision are the three aspects of valid inference [5]. Currently, a large number of methods have been available to objectively and efficiently estimate model parameters and their precision. With the data collected from subjective experiments, correct statistical methods must be adopted to effectively model the underlying principles of a target system. Model specification, or formulation, in its widest sense, is therefore conceptually more difficult than estimating model parameters and their precision [1].

Currently, the analysis of user opinion scores is widely carried out by researchers using parametric statistics. Typical approaches include averaging user scores on each test condition as an absolute mean opinion score (MOS), forming an objective model using a general linear model and verifying the performance using parametric correlation tests.

B. Problem Space

A statistical test is only valid under certain conditions, specified by the requirements and measurements of a model. Therefore, before a particular statistical model is used, the conditions of the model must be verified against the data. For parametric statistical models associated with a normal distribution such as the T-test, which indicates how the mean of two data groups are statistically different from each other, the conditions include [6]:

1) the observations must be independent;
2) the variables must have been measured on at least an interval scale, so that it is possible to interpret the results;
3) the observations must be drawn from normally distributed populations.

Although these crucial conditions for valid statistical analysis have been established for decades by statisticians, they are very rarely verified by researchers in computer science. Several statistical tools such as the Pearson correlation and the Least Square (LS) method have been adopted for model building without solid support of theoretical principles. This absence of statistical validity has become a critical issue, particularly in the field of objective video quality evaluation. This issue is also indicated by international organisations such as the VQEG. Alternative rating scales such as the 11-point scale are adopted over the 5-point scale aiming to amend the distribution of user scores for better statistical validity. However, the MOS metric and parametric statistics are still widely used improperly. The modelling of probability distribution was exploited by Janowski and Nguyen as an alternative of MOS [7], [8]. However, there is still lack of generic and systematic statistical procedures for user score analysis, model selection, parameter estimation, and fitness test.

III. Data Acquisition and Evaluation

In order to systematically introduce the approaches to examine subjective scores and to select appropriate statistical models, we use data from one of our recent subjective experiments as an example. The experiment (firstly introduced in [9]) aims at investigating and modelling users' opinions of the effect of content loss caused by network impairments using content, system and user related metrics. Although this particular test might not share the same objectives as other test plans (such as the ones that evaluate distortions caused by video compression), the testing procedure and data analysis are both generic components of subjective experiments conducted using the same guidance provided by ITU and VQEG [2], [4].

A. Subjective Experiment

1) Test Configuration: Twenty uncompressed video sequences, referred to as the Source Reference Circuit (SRC), with different motion and complexity levels are encoded using the main profile of the H.264 codec. The slice mode is specified as "an entire row of macroblocks per slice". We then use a packet loss emulator to remove the corresponding content from encoded sequence to emulate transmission losses. The loss emulations are applied to all three frame types (i.e. I, P and B). P-frames of three different levels of dependency in GoP are included to investigate the influence of the temporal duration of distortions. Table I shows the frames under test with their dependency in a GoP.

TABLE I
Frames under test

GoP	**I** B B **P1** B B P **B** B P B B **P2** B B P B B P B B **P3** B B

In practice a content packet may contain a varied amount of video content. Therefore the loss of a single packet can lead to distortion with different spatial coverage on video frames. Three sizes of loss are investigated on each frame type. Content loss of one, two and four reference units (Figure 1) are chosen for I and P frames. On B frames where coverage of packet loss is commonly larger due to the more efficient coding mechanism, loss of two, four and eight reference units (Figure 1) are applied.

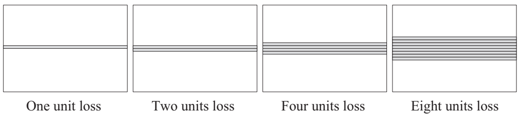


Fig. 1. Spatial coverage of the content loss

As a whole, 600 test conditions with content loss configured and 20 reference conditions with no content loss configured are defined. Each test condition is also referred to as a Hypothetical Reference Circuit (HRC). For instance, "HRC 12_I_2" indicates the test condition of 2 units of content loss experienced on an I frame of the source content SRC 12. HRCs are applied on SRCs to generate test sequences for user tests.

2) Test Procedure: 60 participants took part in our experiment. A Samsung 40-inch Full HD LCD TV is employed as the display device. The purpose-designed testing software AcrVQWin [10] is used for presenting test sequences and collecting user ratings. User opinions were collected using the absolute category rating (ACR) [3] method. ACR is a single-stimulus method with which test sequences are presented individually without being paired with corresponding reference sequences. Participants are asked to provide a quality rating with regard to the visual impact of distortions using the categorical rating shown in Table II. A number of rating scales, including *impairment scale* (Imperceptible; Perceptible but not annoying; Slightly annoying; Annoying; Very annoying) and *quality scale* (Excellent; Good; Fair; Poor; Bad) are available for specific test plans. Because the goal of the experiment is to investigate the perception of impairments, the *impairment scale* is the most suitable rating scale to apply to the ACR method.

B. Overview of Results

Users' scores on all HRCs are presented in the form of a frequency histogram, which gives the distribution of participants' scores on each HRC using the five-point scale. Figure 2 and Figure 3 shows results for SRC 12 and SRC 19.

TABLE II
CATEGORICAL RATING SCALE

Rate the impact of error to the video quality	
Score	Description
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

The differences between users' scores on HRC 12 and HRC 19 illustrate the impact of content characteristics. There is a also clear difference between results on different frame types.

Taking the results for B frames as an example, the content loss of 2 units (i.e. HRC 12_B_2 and HRC 19_B_2) is hardly noticed by participants. When the content loss increases to 8 units, 14 (out of 30) participants noticed the distortion in HRC 19_B_8. Out of the 23 participants that confirmed the perception of an error in HRC 12_B_8, 10 believe the error is "slightly annoying" or even "annoying".

Compared with the results of B frames, content loss in I frames are extremely detrimental, especially on the HRC 12. With merely 1 unit of content loss, 73 percent of participants perceive the impairments in HRC 19_I_1 while 36 percent of them feel the error "slightly annoying" or "annoying". The errors in HRC 12_I_1 are nearly visible to every participant of the experiment while half the participants believe the error "annoying" or "very annoying". With more units of loss introduced, the visibility of corruption increases greatly. On 4 units of loss (i.e. HRC 12_I_4 and HRC 19_I_4), nearly all participants noticed the error. Looking at the distribution of scores it can be observed that SRC 12 is much more vulnerable to I frame impairments than SRC 19. Results for all other test sequences also show clearly shifts of data distribution.

C. Analysis of Experiment Results

Before any statistics are applied, we verify the conditions of widely adopted parametric tests (as summarised in section II-B) against the characteristics of user opinion scores collected from our experiment.

1) Independency of Observation: Every score rated by users in the experiment are independently collected. Each test session is completed by only one participant and the rating process on the HRCs are also independent from each other. Hence, *condition 1* (independent observation) is met.

2) Measurement Scale of Data: For data with interval scales, the categories are ordered and numerical labels or scores are attached. The scores are treated as category averages, medians or mid-points. Differences between scores are therefore interpreted as a measure of separation of the categories [11].

In our experiment, the scale of 1 to 5 attached to each opinion category is arbitrary. Clearly, this numeric representation of each category should therefore not be used to interpret the intensity of any category or the difference between response categories. Furthermore, the psychological distance between the categories should not be considered as equal. For instance, the difference between "imperceptible" and "perceptible" is not equal to the difference between "annoying" and "very annoying" in the psychological scale. For a nominal categorical scale, numbers or other symbols associated with options are used merely for classification.

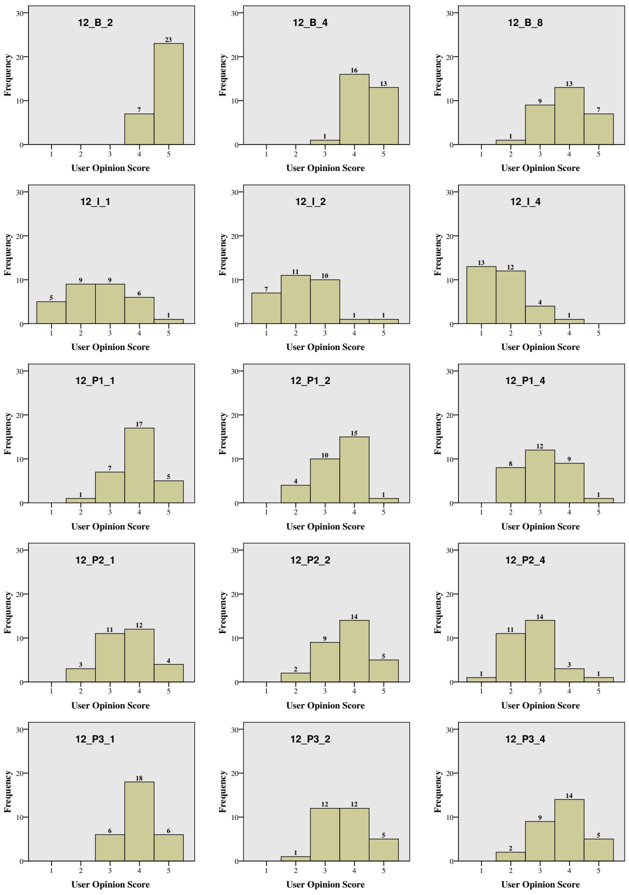


Fig. 2. Results for HRC 12

Furthermore, the five point scale used by the experiment is ordinal. This means that each point on the scale is different from the others (e.g. imperceptible $\neq$ perceptible), but also related (e.g. "imperceptible" $<$ "perceptible") [6]. The five point scale shows a clear ordering of the response categories with respect to the observed perceptual impact of the content loss: "imperceptible" $<$ "perceptible" $<$ "slightly annoying" $<$ "annoying" $<$ "very annoying".

As a whole, the user opinion scores collected in the test are in an ordinal scale. Thus, the test results must not be interpreted using the assigned numerical labels. Parametric statistics are therefore not suitable for data analysis when adequate psychological representations in interval scales are not available. This conclusion also applies to the tests using a *quality scale* (Excellent; Good; Fair; Poor; Bad). Consequently, *condition 2* is not met.

3) Normality of Data Distribution: The normality of the distribution of users' scores on each HRC is a prerequisite

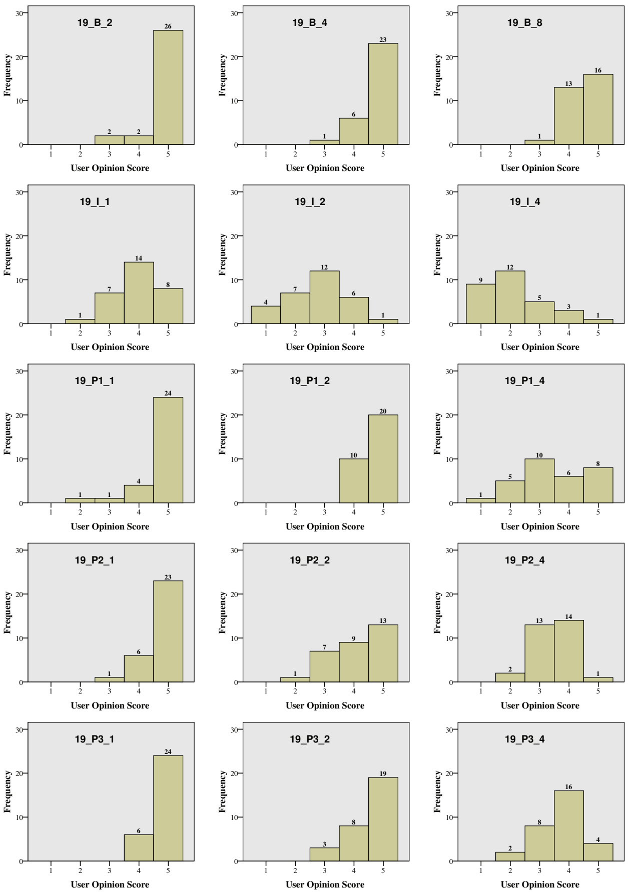


Fig. 3. Results for HRC 19

for statistics based upon a MOS. We use three exclusive normality test tools, the "Skewness test", the "Kurtosis test" and the "Shapiro-Wilk test", to determine whether or not it is statistically legitimate to assume a normal distribution on the test data for parametric modelling.

Skewness measures the degree of symmetry of a probability distribution [12]. If skewness is greater than zero, the distribution is skewed to the right, having more observations on the left. Kurtosis measures the thinness of tails or "peakedness" of a probability distribution [12]. The Shapiro-Wilk test checks the normality assumption by constructing the W statistic, which is the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance [13].

Table III gives results of the Skewness, the Kurtosis and the Shapiro-Wilk test on HRC 12 and HRC 19 as examples. It is concluded from the results of all three tests that the normality of the distribution of the users' scores in subjective experiment cannot be assumed. The same conclusion is also drawn on other HRCs. This means *condition 3* is not met and therefore that statistical analysis methods based upon the assumption of given data being normally distributed, must not be utilised for experiment.

TABLE III
TEST OF NORMALITY

HRC	Skewness	Kurtosis	Shapiro-Wilk	
			Statistic	Significance
12_I_1_1.avi	0.142	-0.849	0.88	0.048
12_I_2_1.avi	0.433	-0.669	0.798	0.003
12_I_4_1.avi	1.649	3.923	0.713	0
12_P1_1_1.avi	-0.415	0.38	0.865	0.028
12_P1_2_1.avi	-0.128	-1.348	0.817	0.006
12_P1_4_1.avi	0.433	-0.669	0.798	0.003
19_I_1_2.avi	0	-0.179	0.799	0.004
19_I_2_2.avi	-0.078	-1.328	0.87	0.034
19_I_4_2.avi	0.715	-0.756	0.848	0.016
19_P1_1_2.avi	-3.326	11.391	0.394	0
19_P1_2_2.avi	-2.405	4.349	0.413	0
19_P1_4_2.avi	-0.124	-0.654	0.896	0.082

4) Summary: This subsection has exploited statistical tools to examine the five-scale user opinion scores against the conditions of the parametric tests that have been commonly adopted to model user ratings. The data analysis shows that the validity of some commonly used subjective QoE metrics must be statistically verified for valid data modelling. For instance, the mean opinion score (MOS) is the *defacto* standard metric for representing user opinions in subjective experiments [14]. Its accuracy, however, cannot be verified because the measurement scale is based on five-point ordinal response categories and because the non-normality of data, representing users' opinions by their arithmetic mean, is statistically incorrect. Consequently, any model that is built referring to MOS is not able to provide valid and interpretable results. Alternative statistical methodologies must therefore be established for the modelling of user opinion scores.

IV. Non-parametric Statistics

Although it is possible to conduct a parametric statistical test for data of any type, the validity and interpretability of the test depends on how the numerical values reflect the underlying principles. For instance the "5" associated with "imperceptible" and the "1" associated with "very annoying" must not be considered as a valid representation of user opinion unless they are verified as being in a correct psychological scale. Non-parametric statistics focus on the order or ranking of scores, not on associated artificial numerical values. Whereas, a parametric test may focus on the difference between the means of two populations, the non-parametric test may focus on the difference between the medians [6]. The advantages of non-parametric statistics are summarised in [6] as follows;

1) If the sample size is very small, there may be no alternative to using a non-parametric statistical test unless the nature of the population distribution is known exactly;
2) Non-parametric statistics typically make fewer assumptions about the data and may be more relevant to a particular situation and research investigation;
3) Non-parametric statistics are available to analyse data which is inherently in ranks as well as data whose seemingly numerical scores has the strength of ranks.

Therefore, non-parametric statistics provide valid analysis tools for the subjective scores rated by human participants on an ordinal scale. Further, the modelling of user scores with parametric statistics requires a relatively large sample size. This requirement can greatly affect the efficiency of time-consuming and costly experiments. Non-parametric statistics, which have much more relaxed sample size requirements, are therefore far more suitable for subjective experiments of a limited scale. This section now presents background and instance of generalised linear models as such a non-parametric statistics for the analysis and modelling of user opinion scores.

A. Generalised Linear Models

The generalised linear model (GLM) generalises the ordinary linear model to encompass non-normal response distributions and modelling functions of the mean. The generalisation enables a more effective modelling of the five-point scale user opinion scores with relaxed requirements on the normality and scale of experiment data.

The generalisation from the ordinary linear models is comprised of three components [15]:

1) The random component of a GLM consists of a response variable with independent observations from a distribution in the exponential family. The raw data as the participants' scores from subjective experiments is taken as the random component of GLM.
2) The systematic component of a GLM relates a vector (η_1,. . .,η_N) to the explanatory variables (for example, the predictor of a video quality model) through a linear model. The covariates x_{i1},. . .,x_{ip} produce the vector given by:
$$\eta_i = \sum_{1}^{p} x_{ij}\beta_j; i = 1, \ldots, N \tag{1}$$
3) The third component is the link function g which may become any monotonic differentiable function.
$$\eta_i = g(\mu_i) \tag{2}$$

An effective link function must be selected based on specifics of the target model in conjunction with observations from experiments. Probit and logit models are effective when there are only gradual changes in cumulative probability, otherwise other link functions should be considered. In particular, linear models using the logit scale or the complementary log-log scale, are found to work well in practice [16]. The changes of cumulative probability for most of the HRCs are gradual in our experiment. Therefore, the logit function is specified as the link function of the model.

B. Cumulative Logit Model

For a five-point user opinion scale, the choice and definition of response categories (opinion scale) is either arbitrary or subjective. It is essential that the nature of the modelling of users' responses should not be affected by the number or choice of response categories. Such considerations lead to modelling the dependence of the response on the independent variables by means of the cumulative response probability as a realisation of GLM (Equation 3) [17]. $\pi_j(x)$ is the probability that score j is rated by human users. $logit[\gamma_j]$ models the logit that users score less or equal to score j. J specifies the total number of response categories (5 in our experiment). A number of $J-1$ logits are established since $P(Y \leq J|x)$ is always 1.

$$logit[\gamma_j] = \log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} = log \frac{\pi_1(x) + \ldots + \pi_j(x)}{\pi_{j+1}(x) + \ldots + \pi_J(x)}, j = 1, \ldots, J-1 \tag{3}$$

A model that simultaneously uses all cumulative logits is

$$logit[P(Y \leq j|x)] = \alpha_j + \beta' x, j = 1, \ldots, J-1 \tag{4}$$

Each cumulative logit of the model has its own intercept (α_j). The α_j is increasing in j, since P(Y≤j) increases in j for fixed *x*, and the logit is an increasing function of this probability. Due to the nature of the cumulative logit each logit for the model has the same effect β and different α_j.

Following the structural form of the model, the intercept and coefficients of variables in the model are estimated from subjective experimental results with the maximum likelihood (ML) estimation. The likelihood indicates the probability that the model can predict the observed data (e.g., subjective ratings) with the independent variables (e.g., metrics) defined. The ML estimate is the parameter value that maximises this function. This is the parameter value under which the observed data has the highest probability of occurrence [18].

C. Calculation of Confidence Intervals

The interpretation of models often involves the examination of predicted outcomes at specific values of the independent variables with confidence intervals (CI) [19]. The confidence intervals give a value range within which the population value falls. Confidence intervals are a standard way of expressing the statistical accuracy of the prediction.

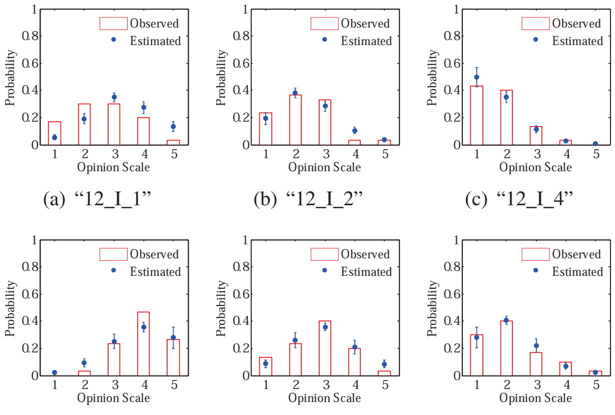
Calculation of CI for the models that are based on the cumulative logit model is not as straight-forward as for parametric models due to the complex link functions. The delta method is a general approach for computing confidence intervals for functions of maximum likelihood estimates. The delta method takes a function that is too complex for analytical computation of the variance and creates a linear approximation of that function. The variance of the simpler linear function is then used for constructing the confidence interval [20]. More details and implementation of the delta method can be found in [21].

D. Calculation of Goodness of Model Fit

When fitting a statistical model, the value of the dependent variable (such as the user opinion score) is considered to be composed of two parts: the systematic component and the error component [22]. The systematic component is a mathematical function of the independent variables that characterise the given observed data among subjects with the independent variables of the model. Although the fitness of the derived model can be roughly examined by comparing the model output to the observed data, the quantitative evaluation of the model fit is impossible with only the systematic component. The error component represents how much the model's output differs from the observed data. The process of examining the values for the error component is referred to as assessing the goodness of fit of the model. In practice the goodness of fit provides crucial information on how the model's output resembles the observed data from the experiments. It is also a useful tool for examining the effectiveness of the model design. Pearson χ^2 and Deviance goodness of fit statistics are two commonly used methods to test whether the observed data are well described by the fitted model.

E. Use Case

For the subjective experiment introduced in section III-A, we use the presented cumulative logit model to form the structure of model, which aims at estimating the distribution of user opinions on the 5-point scale using a number of predictors. Parameters of the model are derived using maximum likelihood estimation. The goodness of model fit and confidence intervals are calculated using statistics introduced in Section IV-C and IV-D. Figure 4 demonstrates the results of the estimation with 95 % confidence interval for HRC 12_I and HRC 19_I. It clearly demonstrate how objective models based on cumulative logit statistics can capture the distribution data for different test conditions. Furthermore, the distribution format provides more interpretable insights of users' opinion which are not possible by MOS.



(a) "12_I_1" (b) "12_I_2" (c) "12_I_4"

(d) "19_I_1" (e) "19_I_2" (f) "19_I_4"

Fig. 4. Model estimation results of HRC 12_I and HRC 19_I

V. Conclusion

Subjective user opinions derived from experiments are commonly exploited to construct objective quality models using statistical analysis. Although a number of parametric statistics have been widely adopted for this task, the conditions for using parametric statistics are rarely verified. Our tests verify that non-normally distributed user opinion scores in nominal or ordinal responses should not be analysed using parametric statistics. We therefore introduced methods and tools for valid model selection and parameter estimation. Finally, a group of test result demonstrated the effectiveness and advantages of the user opinion distribution over the MOS.

VI. Acknowledgements

The work presented in this paper is supported by European Commission within the FP7 Project: P2P-Next and FIRM project (Framework for Innovation and Research in MediaCityUK).

References

[1] K. Burnham and D. Anderson, *Model selection and multimodel inference: a practical-theoretic approach*. Springer, New York, 2002.

[2] "Methodology for the subjective assessment of the quality of television pictures," *ITU-R Recommendation BT.500-11*, 2002, iTU.

[3] "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, 1996, iTU.

[4] "Report on the validation of video quality models for high definition video content," *Video Quality Experts Group*, 2010, http://www.its.bldrdoc.gov/vqeg/.

[5] R. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, pp. 309–368, 1922, royal Society of London.

[6] S. Siegel and N. Castellan, *Nonparametric statistics for the behavioral sciences, Second edition*. McGraw-Hill, 1988.

[7] L. Janowski and Z. Papir, "Modeling subjective tests of quality of experience with a generalized linear model," in *Proc. Int. Workshop Quality of Multimedia Experience QoMEx 2009*, 2009, pp. 35–40.

[8] R. H. Le Thu Nguyen and J. Jusak, "A pilot study to assess quality of experience based on varying network parameters and user behaviour," *International Proceedings of Computer Science and Information Technology*, vol. 5, p. 30, 2010.

[9] M. Mu, A. Mauthe, R. Haley, and F. Garcia, "Discrete quality assessment in IPTV content distribution networks," *Elsevier Journal of Signal Processing: Image Communication*, 2011, elsevier.

[10] "AcrVQWin," *http://www.acreo.se/*, 2007.

[11] P. McCullagh and J. Nelder, *Generalized linear models*. Chapman & Hall/CRC, 1989.

[12] H. M. Park, "Univariate analysis and normality test using SAS, STATA, and SPSS." The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University, Tech. Rep., 2008, available at http://www.indiana.edu/ statmath/stat/all/normality/normality.pdf (accessed on 20/07/2011).

[13] S. Shapiro and M. Wilk, "An analysis of variance test for normality (complete samples)," *Journal of Biometrika*, vol. 52, no. 3-4, p. 591, 1965, oxford University Press.

[14] "Final report of VQEG's Multimedia phase I validation test," *Video Quality Experts Group*, 2008, http://www.its.bldrdoc.gov/vqeg/.

[15] A. Agresti, *Categorical data analysis, second edition*. John Wiley & Sons, Ltd,, 2002.

[16] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980, blackwell Publishing.

[17] *SPSS Base 16.0 users guide*. Pearson Education, 2007.

[18] A. O'Connell, *Logistic regression models for ordinal response variables*. Sage Publications, Inc, 2006.

[19] R. Fisher, *Statistical methods and scientific inference*. Oliver and Boyd London, 1956.

[20] J. Long, J. Freese, and L. StataCorp, *Regression models for categorical dependent variables using STATA*. STATA PRESS, 2006.

[21] J. Xu and J. Long, "Confidence intervals for predicted outcomes in regression models for categorical outcomes," *Stata Press: Stata Journal*, vol. 5, no. 4, p. 537, 2005.

[22] D. Hosmer, S. Taber, and S. Lemeshow, "The importance of assessing the fit of logistic regression models: A case study," *American Journal of Public Health*, vol. 81, no. 12, p. 1630, 1991, american Public Health Association.