Compensating for Type-I Errors in Video Quality Assessment

Kjell Brunnström, Samira Tavakoli and Jacob Sogaard

Abstract—This paper analyzes the impact on compensating for Type-I errors in video quality assessment. A Type-I error is to incorrectly conclude that there is an effect. The risk increases with the number of comparisons that are performed in statistical tests. Type-I errors are an issue often neglected in Quality of Experience and video quality assessment analysis. Examples are given for the analysis of subjective experiments and the evaluation of objective metrics by correlation.

Keywords—Type-I error, video quality, statistical significance, Student T-test, Bonferroni

I. INTRODUCTION

Currently, subjective experiments are the best way to investigate the user’s Quality of Experience (QoE) for video. Typically, in such experiments, panels of observers rate the quality of video clips that have been degraded in various ways. When analyzing the results, the experimenter often computes the mean over the experimental observations, a.k.a. the Mean Opinion Scores (MOS) and applies statistical hypothesis tests to draw statistical conclusions. A statistical hypothesis test is done by forming a null hypothesis (H_0) [1] and an alternative hypothesis (H_1) that can be tested against each other. For example, in video quality assessment, often the hypothesis test will have the null hypothesis, H_0, that the two underlying MOS values are the same and the alternative hypothesis, H_1, that they are different. If the result is significant, the experimenter knows with high probability (typically 95%) that H_1 is true and in this case, that the MOS values are different. However, there is still a small risk (5% in this case) that this observation is only by chance. This is a Type-I error—to incorrectly conclude H_1 is true when in reality H_0 is true.

When there are more pairs of MOS values to compare, each comparison has the above mentioned small risk of error. This risk of an error increases with the number of comparisons and can be estimated by: 1 - (1 - a)^n, where a is the confidence level per comparison and n is the number of comparisons [1]. For 100 comparisons at a 95% confidence level, this equals more than a 99% risk of at least one Type-I error.

In this paper, we demonstrate the consequences of Type-I errors in video quality assessment. The work was motivated by a recent study [2], where in spite observing large absolute differences between MOS values, no statistical significance was observed. There are also important discussions when to use parametric or non-parametric statistical methods and if normal distribution assumptions are valid or not in video quality assessment, but those are outside the scope of this paper. Furthermore, there is a difficult trade-off while securing against Type-I errors, which increases the risk of committing Type-II errors (i.e. not finding an effect while it is there). But we focus on the Type-I error here, since we feel that this is more often neglected.

II. METHOD

There are various statistical methods to compensate for Type-I errors. It is important to distinguish between planned comparison and post-hoc testing. If a set of comparisons are planned before the data is collected, then n effectively drops. That is, n is the actual number of comparisons planned ahead instead [1]. Otherwise all possible comparisons should be taken into account.

A common way to compare a set of means is to perform an Analysis of Variance (ANOVA) followed by a post-hoc test. This is a two step approach where first ANOVA indicates whether there is an overall effect, then a more refined tests (such as Tukey HSD) analyzes whether there are any pairwise significant differences. However, it is quite difficult to estimate how big of an influence a particular number of comparisons has on the efficiency of the statistical test. Fortunately, there is also a rather straightforward method, suggested by Bonferroni [1], where the considered significance level (a) is divided with the number of comparisons (n) so that the significance level for each comparison will be a/n. The advantage here is that it can be combined with simple tests like the Student’s T-test. The disadvantage is that it can be overly conservative.

In this study, we consider the influence of multiple comparisons on the number of test subjects required and on the differences between MOS that are statistically significant. We also consider the performance evaluation of objective metrics, based in ITU-T Rec. P.1401 [3]. To this end, we analyze Pearson’s correlation for multiple comparisons.

To analyze an effect, we assume the Student’s T-test with equal standard deviations and the same number of data points in the two mean values, based on independent data samples. This gives the simplified formula t_{obs} = \frac{n \cdot \mu_1 - \mu_2}{\sqrt{\sigma^2}}. The degrees of freedom are (2n-2). For certain values of the difference between the means (\mu_1-\mu_2), the number of data points (n) and the standard deviations (\sigma), we can calculate the probability of
III. RESULTS

Fig. 1. Top, middle: Probability of significance for subjective experiments. ‘Alpha’ and ‘diff’ denote the confidence level per comparison and MOS difference in order. Bottom: Probability of significance for Pearson correlations with a difference of 0.05, where N is the number of data points.

We can analyze the requirements for getting statistical significance by calculating the p for different input values. This simplification is not directly useful for most video quality experiments. However, our simplification covers the important case where an experiment has been repeated by different labs or different panels of observers. For instance, when comparing two experiments using the same distorted videos, the experimenter might want to test whether the MOS difference is 1.0 or more on a 5-level scale (e.g. in one lab a video is rated “good”, but at another it is just rated “fair”).

IV. CONCLUSIONS

In this paper, we investigated the effect of multiple comparisons on the statistical level of significance that can be expected in subjective studies and objective metrics evaluations. This effect can result in the Type-I error, which is often neglected and therefore leads to wrong conclusions. Our results show that there could be arguments to increase the number of test subjects normally used according to standardized recommendations—especially, if the goal is to detect a 1.0 MOS difference. Further, for objective metric comparisons using correlation coefficients, it is difficult to find any significance with few data points and correlations below 0.9. In this case, multiple comparisons have a large impact on the final conclusion that can be drawn.

REFERENCES