# Data Analysis in Multimedia Quality Assessment: Revisiting the Statistical Tests

Manish Narwaria, Lukáš Krasula, and Patrick Le Callet

***Abstract*—Assessment of multimedia quality relies heavily on subjective assessment, and is typically done by human subjects in the form of preferences or continuous ratings. Such data is crucial for analysis of different multimedia processing algorithms as well as validation of objective (computational) methods for the said purpose. To that end, statistical testing provides a theoretical framework towards drawing meaningful inferences, and making well grounded conclusions and recommendations. While parametric tests (such as $t$ test, ANOVA, and error estimates like confidence intervals) are popular and widely used in the community, there appears to be a certain degree of confusion in the application of such tests. Specifically, the assumption of normality and homogeneity of variance is often not well understood. Therefore, the main goal of this paper is to revisit them from a theoretical perspective and in the process provide useful insights into their practical implications. Experimental results on both simulated and real data are presented to support the arguments made. A software implementing the said recommendations is also made publicly available, in order to achieve the goal of reproducible research.**

## I. Introduction

The growth of low-cost devices has virtually made multimedia signals an integral part of our daily lives. Todays end users are constantly interacting with multimedia, and are more demanding in terms of their multimedia experience, and perceptual quality is one of the intrinsic factors affecting such interaction. As a result, assessment of perceptual quality is an important aspect in todays multimedia communication systems [1]. The most reliable way of quality estimation typically involves the use of a human subject panel who provides ratings/preferences for the targeted multimedia content [1], [2]. This is referred to as subjective assessment. In contrast, objective estimation of quality relies on the use of computational (mathematical) models [3] that are expected to mimic subjective perception.

Parametric statistical tests find extensive application in multimedia quality estimation mainly for two purposes. First, they are used to compare and analyze subjective data collected from human participants. For instance, a $t$-test can be used to compare Mean Opinion Score (MOS) from two different conditions in a variety of applications (eg. analyzing codec performance [4], investigating the effect of upscalers on video quality [5], studying optimization criteria in HDR tone mapping [6] and so on). Analysis of Variance (ANOVA) is also a commonly used technique for analyzing the effect of two or more factors/treatment levels and their interactions. These include identifying audiovisual interactions [7], examining the impact of reflections in HDR video tone mapping [8], investigating the effect of resolution, bit rate and color space on under water videos [9], studying the possible impact of compression level and type of content on perceptual quality towards finding optimal presentation duration in subjective quality assessment [10] etc. Second, these tests are used to validate objective (computational) methods against subjective data. This can in turn be used to statistically compare several objective methods in terms of their prediction accuracies as compared to the subjective data. Such validation studies are obviously central to benchmarking objective methods before they can be deployed in practice.

The need for statistical testing arises due to the fact that subjective studies use a finite sample of human subjects. Therefore, these tests can help in generalizing and making inferences for the population. For that purpose, parametric tests such as $t$-test, $F$-test, ANOVA, and error estimation (eg. using confidence intervals) are widely used in the community. While the application of parametric tests is generally straightforward (aided by the availability of numerous software packages), the interpretation of the results requires some care. In particular, statistical tests in many cases are simply treated as *black boxes*, and are applied without considering the practical implications of the assumptions in these tests.

As the name implies, such tests are based on *apriori* knowledge of parameterizable probability distribution functions (eg. $t$ distribution, $F$ distribution which are respectively characterized by one and two degrees of freedom.). While it is true that parametric tests are distribution dependent (as opposed to non-parametric tests which are some times referred to as being *distribution-free*), there appears to be some confusion regarding the assumptions made in these tests. In particular, the assumption of normality and homogeneity of variance in many cases appears to be not well understood for both subjective and objective data analysis. In practice, these assumptions are sometimes considered as bottle necks in applying parametric tests. As a result, nonparametric tests are recommended if the data violates one or both the assumptions. A typical approach to applying parametric statistical tests is depicted in the left flow diagram in Figure 1, and consists of arriving at one of the three decisions $D_1$, $D_2$ or $D_3$:

- $D_1$: normality checks (eg. JB test, K-S test) are applied to examine if the given subjective/objective data is normal. If such normality checks determine the data to be *nonnormal* then nonparametric tests are carried out.
- $D_2$: If the normality test determines the data to be *normal*, then homogeneity of variance is tested by applying a

Manish Narwaria is with Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, 382007, India. Lukáš Krasula and Patrick Le Callet are with LS2N/IPI group, University of Nantes, 44306, France e-mail: (manish_narwaria@daiict.ac.in, lukas.krasula@univ-nantes.fr, patrick.lecallet@univ-nantes.fr).

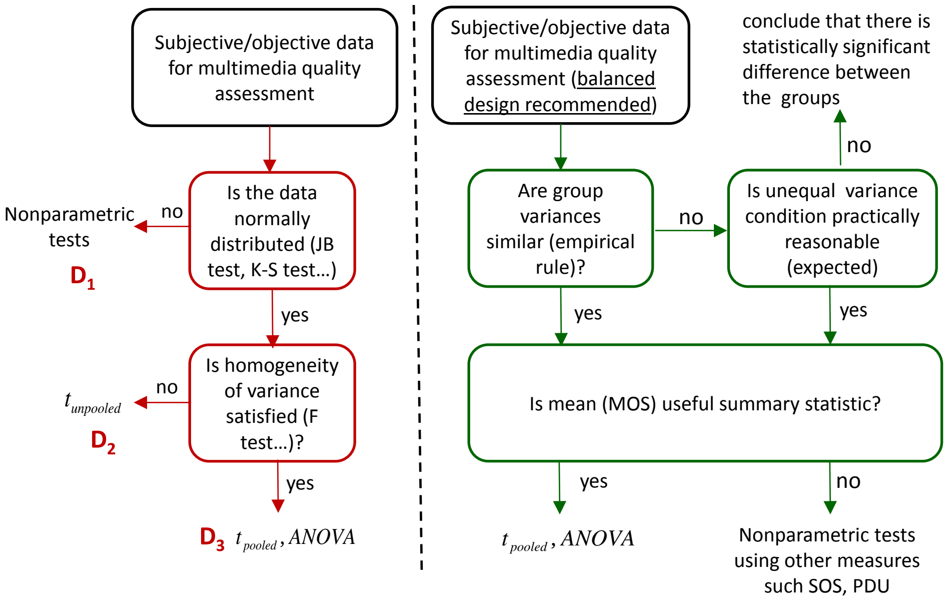

Subjective/objective data for multimedia quality assessment
Is the data normally distributed (JB test, K-S test...)
no
Nonparametric tests
D1
yes
Is homogeneity of variance satisfied (F test...)?
no
t unpooled
D2
yes
D3 t pooled , ANOVA
Subjective/objective data for multimedia quality assessment (balanced design recommended)
Are group variances similar (empirical rule)?
no
Is unequal variance condition practically reasonable (expected)
no
conclude that there is statistically significant difference between the groups
yes
yes
Is mean (MOS) useful summary statistic?
yes
t pooled , ANOVA
no
Nonparametric tests using other measures such SOS, PDU

Fig. 1: Typical procedure of applying parametric tests (the left flow chart) and the recommended approach (right flow diagram). The drawbacks associated with making decisions $D_1$, $D_2$ or $D_3$ are discussed in sections II and III. Figure best viewed in color.

test of variance (eg. Levene's test, $F$ test etc). If the groups/samples do not satisfy the said assumption then modified tests (eg. unpooled $t$ test) are applied which do not use pooled variance in computing the test statistic.

- $D_3$: If the data satisfies both assumptions of normality and homogeneity of variance then the usual $t$ test or ANOVA (which employ pooled variance) are applied.

In this paper, we seek to draw attention to few drawbacks associated with such decisions. Specifically, we revisit theoretical formulations and the resultant practical implications to highlight shortcomings and recommend alternative approach (right flow diagram in Figure 1) in the light of the said assumptions. We emphasize that these assumptions should not be viewed as constraints or bottle necks in the application of parametric tests. Instead these should be carefully considered and understood in the context of their practical implications. Subsequently, we provide a set of recommendations to ameliorate some of the drawbacks that may stem from either wrong interpretation or application of the said assumptions in parametric testing. A software implementing the said recommendations is also made publicly available[*], in order to achieve the goal of reproducible research.

The remainder of the paper is organized as follows. In section II we analyze the distributional assumptions in parametric test. Section III provides an analysis of the assumption of homogeneity of variance. Section IV points out the practical implications in the context of multimedia quality assessment. In section V we present the experimental results and analysis while Section VI lists a set of recommendations towards proper use of parametric testing in the context of the said assumptions. We provide concluding thoughts in section VII.

## II. Revisiting Distributional Assumptions in Parametric Tests

Parametric tests require certain assumptions including the assumption of normality, homogeneity of variance and data independence. As highlighted in left flow diagram in Figure 1, normality checks have usually been applied on subjective or objective data [2], [3], [4], [11]. Such use of normality checks indicates that the assumption of normality is, in many cases, misunderstood to be applicable on the data for which statistical tests are to be carried out. This is, however, incorrect in the light of the fact that all parametric tests essentially work by locating the observed test statistic on a known probability distribution function. Then, depending on the desired significance level and the location of test statistic, one typically accepts or rejects the null hypothesis. For example, in $t$-test, the $t$-statistic is first computed from the observed sample. This $t$-statistic is then compared with values from a $t$-distribution (corresponding to the particular degrees of freedom). In other words, the computed test statistic ($t$-statistic, $F$-statistic etc.) is assumed to follow the corresponding distribution ($t$-distribution in $t$-test, $F$-distribution in $F$-test and ANOVA etc.).

Thus, the more appropriate question to be asked in parametric testing is whether the test statistic follows the assumed distribution (rather than the data being normally distributed). The answer to such question requires that the subjective (or objective) test be repeated for a large number of times, each time using a different sample (both in terms of human subjects and content). Then, in each instance, the test statistic can be

[*] https://sites.google.com/site/narwariam/home/research

computed to obtain its sampling distribution. This process is, however, neither practical for obvious reasons nor desirable. Instead, one can rely on the fundamental central limit theorem (CLT). Informally, the CLT states that the sampling distribution of the arithmetic mean (and sum) will approach a normal distribution as the sample size increases, regardless of the underlying population distribution [12]. It is due to this result that the test statistic in parametric tests are guaranteed to follow the assumed distribution, provided that the sample size is large enough (approaching infinity in theory).

We begin by considering two populations $\mathbf{p_1}$ and $\mathbf{p_2}$ with means $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively. In the context of multimedia quality assessment, these populations will typically represent the collection of subjective (or objective) opinion scores for two conditions (eg. subjective or objective quality scores for two profiles of a video codec, indivudual quality scores for audiovisual content corresponding to two parameter settings, quality scores for content rendered by two depth image based rendering methods, individual quality scores for two tonemapped HDR videos and so on) for which we need to compare mean quality scores i.e. $\mu_1$ and $\mu_2$. Assume that $\mathbf{p_1}$ and $\mathbf{p_2}$ are sampled i.e. subjective or objective assessment is actually performed on a set of content using a sample of human subjects or using objective methods. Let the corresponding samples be denoted by $\mathbf{x_1} = [x_{11}, ..., x_{1n_1}]$ and $\mathbf{x_2} = [x_{21}, ..., x_{2n_2}]$ where $n_1$ and $n_2$ are the sample sizes, and the sample observations are assumed to be independent and identically distributed (iid) random variables. Note that there are no assumptions regarding the distribution of either the populations ($\mathbf{p_1}$ and $\mathbf{p_2}$) or corresponding samples ($\mathbf{x_1}$ and $\mathbf{x_2}$).

### *A. Sampling distribution of test statistic in t-test*

Let $\overline{x}_1$, $\overline{x}_2$ and $s_1^2$, $s_2^2$ denote the sample means and variances, respectively. Then the goal of the analysis is to infer if $\mu_1 = \mu_2$ (the *null* hypothesis) or not. To that end, one can employ the $t$-test. To define the $t$-statistic, we use the result from the CLT i.e.

$$\overline{x}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \text{ and } \overline{x}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right) \tag{1}$$

Then, the difference between the samples means will also be normally distributed i.e.

$$\overline{x}_1 - \overline{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \tag{2}$$

By standardization, we have

$$\frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \tag{3}$$

Note that in eq. (3) only the numerator is a random variable while the denominator is constant. However, in practice, the population variance is generally not known. We therefore need to use sample variance as an *unbiased estimator* of the population variance. To proceed further, we consider two cases for defining the *null* hypothesis.

*1) Case 1: Samples drawn from same population:* We can define the *null* hypothesis as $H_0$ : the two samples are taken from the same population. This implies that not only are we assuming the population means to be equal but other population parameters including variances are equal. Thus, we have $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say). In order to obtain a more accurate estimate of the (common) population variance, we can employ the pooled variance $s_p^2$ which is defined as

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} \tag{4}$$

Thus, under $H_0$, the denominator in eq. (3) can be modified accordingly and the $t$-statistic defined as

$$t_{pooled} = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, df_{pooled} = n_1 + n_2 - 2 \tag{5}$$

With the said modification, the reader will now note that the denominator in eq. (5) is also a random variable, unlike eq. (3) where it was a constant. Thus, $t_{pooled}$ is a ratio of two random variables. The numerator is the difference of two independent normally distributed random variables ($\overline{x}_1$ and $\overline{x}_2$), and will therefore be normally distributed [13]. Further, the squared denominator will be equal to $\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}$ which denotes the variance of the said normal distribution in the numerator. Hence, the denominator in eq. (5) will be chi-squared distributed [13]. Accordingly, the test statistic $t_{pooled}$ is characterized by the ratio of normally and square root of chi-squared distributed variables. It will therefore be approximately[†] distributed according to the $t$-distribution [13] with $df_{pooled} = n_1 + n_2 - 2$ degrees of freedom, and this will be irrespective of the distribution of either the populations ($\mathbf{p_1}$ and $\mathbf{p_2}$) or corresponding samples ($\mathbf{x_1}$ and $\mathbf{x_2}$).

*2) Case 2: Samples drawn from two different populations with same population mean:* In the second case, we assume that the two samples have been drawn from two different populations with same population mean i.e. $\mu_1 = \mu_2$ (but $\sigma_1^2 \neq \sigma_2^2$). Hence, other population parameters such as variance or any other statistic need not be equal. Then, we can use sample variances as an estimate of the two population variances, and under the assumption of the *null* hypothesis, eq. (3) can be modified to obtain the following test statistic

$$t_{unpooled} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \; df_{unpooled} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}} \tag{6}$$

In practice, we use $s_1^2$ and $s_2^2$ to compute $df_{unpooled}$ in eq. (6) because $\sigma_1^2$ and $\sigma_2^2$ are not known. We will discuss the two cases in section III.

[†] In theory, the sample size should tend to infinity for the sample means to be normally distributed according to CLT. However, in practice, smaller samples sizes allow us to approximate the assumption of normality, regardless of population or sample distribution.

### B. The case of ANOVA and $F$-test

The sampling distribution of the test statistic ($F$) in $F$-test (ANOVA also relies on $F$-test) is assumed to follow the $F$-distribution [13]. It can be shown that this assumption is valid irrespective of the data distribution with the same caveat concerning the CLT mentioned in the previous sub-section.

Before doing that, we assume that there are $k$ groups each with $n_i$ observations (let the total number of observations be denoted by $M = \sum_{i=1}^{k} n_i$), and define the following: mean $\overline{x}_i$ of $i^{th}$ group, grand mean $\overline{X}$ and variance $s_i^2$ of the $i^{th}$ group as

$$\overline{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, s_i^2 = \frac{\sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i\right)^2}{n_i}, \overline{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{k} n_i} \tag{7}$$

The $F$-statistic in ANOVA is defined as the ratio of inter-group (i.e. between groups) and intra-group (i.e. within each group) variations. We denote these quantities by $SS_B$ and $SS_W$, respectively, with the corresponding degrees of freedom being $df_B$ and $df_W$. Then, the $F$-statistic is computed as

$$F = \frac{SS_B/df_B}{SS_W/df_W} = \frac{\sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{X}\right)^2 / (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i\right)^2 / (M-k)} \tag{8}$$

By noting that the denominator in eq. (8) is essentially a weighted sum of individual group variances, we can view the $F$-statistic as

$$F = \frac{\sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{X}\right)^2 / (k-1)}{\frac{n_1 s_1^2 + n_2 s_2^2 + \ldots + n_k s_k^2}{(n_1-1)+(n_2-1)\ldots+(n_k-1)}} \tag{9}$$

One can see that the numerator in eq. (9) is squared difference of two normally distributed variables ($\overline{x}_i$ and $\overline{X}$), and will be thus chi-squared distributed. The denominator can be seen to be very similar to the pooled variance used in eq. (4), and will be chi-squared distributed following similar arguments. It follows that $F$ is a ratio of two chi-squared distributed random variables which in turn implies that it will be approximately distributed according to the $F$-distribution (with $k-1$ and $M-k$ degrees of freedom). Once again, this is independent of the distribution of the population or the groups, and only relies on the approximations related to sample size as required in the CLT.

### C. Data normality checks: are they required?

As discussed in previous sub-sections, the CLT being a theoretical result only provides asymptotic approximation in that as sample size tends to infinity the sampling distribution of mean tends to be normally distributed, and this holds irrespective of the sample or population distribution [12]. Note that the CLT does not specify any sample size above which the said sampling distribution will be normal. In practice, smaller sample sizes are generally sufficient to allow reasonable approximations. For instance, in the context of subjective quality assessment, Ref. [14] recommends a minimum of 15 subjects while the authors in [15] suggested using at least 24 subjects for audiovisual quality measurement. Because the sampling distribution of mean is directly or indirectly used in computing the test statistics such as $t$, $F$ etc., there are no requirements of normality (or any other distribution) on the data to be analyzed. It is, therefore, not surprising that previous works [2], [16], [17] have noted that parametric tests such as ANOVA are *robust to non-normal data distributions*, and the focus on distributional assumptions in these tests is not required [18].

The second theoretical argument against the application of normality checks before conducting parametric tests is the inflation of Type I error probability. A commonly adopted strategy is to first check whether the given sample/data is normally distributed or not. To that end, normality tests such as the Kolmogorov-Smirnov (K-S) test, Jarque-Bera test, Shapiro-Wilk test etc. are popular. If the tests determine the given data is normally distributed then a parametric test is used. Otherwise, a non-parametric test is performed. As a result of this two-step process, there will be an increase in type I error probability. Assume that $H_0^*$ : given data is normally distributed (the *null* hypothesis in a normality test) and $H_0$ be the *null* hypothesis of the test that will follow. Then, the probability of rejecting $H_0$ can be written as the sum of mutually exclusive events i.e.

$$\begin{aligned} P(\text{reject } H_0) = P(&\text{reject } H_0 \text{ and not reject } H_0^*) \\ +P(&\text{reject } H_0 \text{ and reject } H_0^*) \end{aligned} \tag{10}$$

In the above equation the first expression on right hand side corresponds to the case of using a parametric test while the second expression corresponds to the use of a suitable non-parametric test. Because the critical regions corresponding to the parametric and non-parametric tests will be in general different, the resultant critical region which is a union of the critical regions of the individual tests is increased. Consequently, the probability[‡] to reject $H_0$ (when it is true) is increased thereby increasing the probability of a type I error.

The third argument against the use of normality tests is the theoretical contradiction concerning the sample size. It is known that most normality tests, by definition, tend to reject the *null* hypothesis $H_0^*$ (given data is normally distributed) as the sample size increases. For instance, in the JB test for normality, the test statistic value is directly proportional to the sample size. In other words, larger the sample size, it is more likely to be determined as non-normal. However, according to CLT, the approximation of normality of the sampling distribution of mean improves as the sample size increases. This leads to a contradiction between the requirement of data normality and the asymptotic behavior in the CLT.

[‡] This probability value is not related to the $p$ value of the significance test. Instead, it refers to the probability (over repeated trials) of making a type I error i.e. rejecting $H_0$ when it is true.

While other methods such visual (eg. histogram visualization, normal probability plots) or those based on empirical rules (eg. if sample kurtosis is between 2-4, then the sample is deemed to be normally distributed) can overcome the limitations associated with the more formal normality tests, these are not required because it is the normality of sampling distribution of mean that is needed rather than the data being normal.

## III. To pool or not to pool?

In this section, we analyze the assumption of homogeneity of variance and point out the theoretical aspects that need to be considered in the context of this assumption. The relevant practical considerations will be discussed in the next section.

### *A. Should homogeneity of variance be checked?*

As discussed in the previous section, the *null* hypothesis can be defined in two cases. For Case 1, we require the assumption of homogeneity of variance (i.e. $\sigma_1^2 = \sigma_2^2$) and is applicable in the context of ANOVA (for more than two groups) and $t_{pooled}$ (for two groups). Note that both the tests use an estimate of the pooled variance in order to compute the corresponding test statistic. On the other hand, Case 2 does not require homogeneity of variance and is applicable in defining the test statistic $t_{unpooled}$. Therefore, $t_{unpooled}$ is widely used in statistical data analysis and has been included in many statistical packages such as SPSS. However, it can be noted that in general $df_{unpooled} < df_{pooled}$ (except when $\sigma_1^2 = \sigma_2^2$ and $n_1 = n_2$, in which case both are equal), and hence the use of $t_{unpooled}$ will increase the probability of Type II error (i.e. the test will be more conservative). In light of this, a popular and seemingly logical strategy is to first conduct a preliminary test of variance based on which a decision to either use $t_{pooled}$ (or ANOVA) or $t_{unpooled}$ (if the test of variance leads to the conclusion that $\sigma_1^2 \neq \sigma_2^2$).

Notice that this strategy, however, involves cascaded use of the given data in rejecting or accepting two hypotheses (one from test of variance and the other from the $t$-test). In other words, two significance tests are performed on the same data. As a consequence, the Type I error probability will be increased [13]. Suppose $H_0^{**} : \sigma_1^2 = \sigma_2^2$ (the *null* hypothesis in a preliminary variance test for equality of population variances) and $H_0 : \mu_1 = \mu_2$ be (the *null* hypothesis for the $t$-test that will follow). Then, the probability of rejecting $H_0$ in this case can be written as (similar to eq. 10) the sum of probability of rejecting $H_0$ when $H_0^{**}$ is not rejected and the probability of rejecting $H_0$ when $H_0^{**}$ is also rejected. Following the same arguments as in section II-C, the resultant critical region which is a union of the critical regions of the individual $t$-tests is increased thereby inflating the probability of Type I error.

Further, note from eq. (6) that the degrees of freedom for $t_{unpooled}$ depends on population variances $\sigma_1^2$ and $\sigma_2^2$, and will therefore be a random variable in case these are estimated from sample variances (which is practically the more likely case). As a result, its analysis, both theoretical and experimental is more complicated due to the fact that its distribution is not independent of sample variances [19]. Thus, the interest in $t_{unpooled}$ is more from a theoretical perspective in that it allows for a *correction* in degrees of freedom which in turn renders it valid in cases when population variances are not equal. In practice, however, it is more relevant to consider the implications of comparing means of two populations whose spread (variances) are different. Hence, applying statistical tests for checking homogeneity of variance prior to using $t$ test, ANOVA etc. is not recommended due to theoretical (due to increased probability of type I error) reasons, and is of less interest in practice.

### *B. The case of balanced design*

It can be shown that the test statistic $t_{pooled}$ is valid even if $\sigma_1^2 \neq \sigma_2^2$ provided that the sample sizes are equal (balanced design). To prove this, we compare the distributions of $t_{unpooled}$ and $t_{pooled}$ by writing them in terms of the theoretical $t$ distribution [19] in the following form:

$$t_{pooled} = c_{pooled} \cdot t_{df_{pooled}}, t_{unpooled} = c_{unpooled} \cdot t_{df_{unpooled}} \tag{11}$$

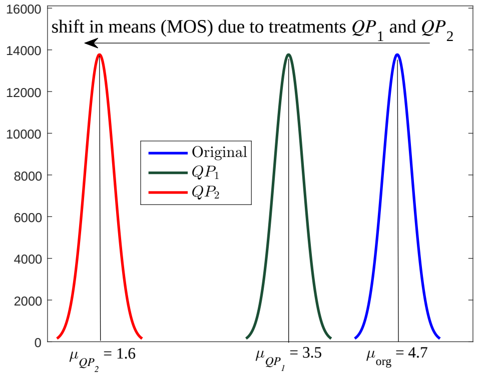
where $t_{df_{pooled}}$ and $t_{df_{unpooled}}$ are the $t$ distributions with respective degrees of freedom. Thus, for $t_{pooled}$ and $t_{unpooled}$ to follow the respective theoretical $t$ distributions the corresponding multiplicative factors $c_{pooled}$ and $c_{unpooled}$ should be equal to 1. It can, however, be shown [19] that while $c_{unpooled}$ is always equal to 1, the value of $c_{pooled}$ depends on sample size and population variances i.e.

$$c_{pooled} = \sqrt{\frac{(n_1 + n_2 - 2)\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}{\left(\frac{1}{n_1} + \frac{1}{n_1}\right)\left\{(n_1 - 1)\,\sigma_1^2 + (n_2 - 1)\,\sigma_2^2\right\}}} \tag{12}$$

From the above equation, it is easy to see that $c_{pooled} = 1$ if the population variances are equal ($\sigma_1^2 = \sigma_2^2$). However, $c_{pooled}$ is also equal to 1 if sample sizes are equal ($n_1 = n_2$). In other words, $t_{pooled}$ will follow the expected theoretical distribution if balanced design is used, despite the violation of the assumption of homogeneity of variance. Because several practical applications tend to target a balanced design i.e. equal sample sizes, the use of $t_{pooled}$ is valid in such cases even if sample variances differ by a large amount. Particularly, in case of multimedia quality assessment, the use of balanced design is common. For instance, typical subjective quality assessment tests use the same number of human subjects to evaluate the quality of different conditions (although the subject panel may or may not comprise of the same subjects in evaluating the quality of each condition).

## IV. Practical considerations in the domain of multimedia quality assessment

In this section, we discuss the assumption of homogeneity of variance from the practical view point, and take an illustrative example from the domain of video quality assessment. Let us consider that an original (i.e. undistorted) video sequence is viewed and rated for its visual quality by all the concerned observers on a scale of 1 (worst) to 5 (excellent). Hence,



(a)

Fig. 2: Illustration of treatment effects $QP_1$ and $QP_2$. The shift in location does not alter the variance of the groups. The values of $\mu_1$, $\mu_2$ and $\mu_3$ are assumed for illustration only. Figure best viewed in color.

this set of individual ratings forms the population of interest $P_{org}$ for this condition (i.e. undistorted video). We can express each element of $P_{org}$ as $P_{org}^{(i)} = \mu_{org} + \epsilon_i$ where $\mu_{org}$ is the mean of $P_{org}$ and $\epsilon_i$ denotes the random error (with zero mean and finite variance) that will be introduced in each individual rating. This error term can be used to take into account the fact that some observers may be more critical (so their corresponding ratings will be less than $\mu_{org}$) while others may be less critical (i.e. their ratings are expected to be higher than $\mu_{org}$) of the video quality. Suppose the said video is now compressed using two quantization parameter ($QP$) values $QP_1$ and $QP_2$ and $QP_2 > QP_1$ ($QP$ is employed in video compression as a measure to quantify quantization levels, higher $QP$ implies higher quantization and in general lower video quality).

### A. The case of systematic treatment effect

In the considered example, quantization can be considered as a treatment that is applied to the original video. Assuming all other conditions to be identical (i.e. same display, ambient light, viewing distance etc.), the treatments $QP_1$ and $QP_2$ will decrease the video quality and essentially cause a shift in means (MOS). In other words, the intervention in original video will result in shifted (in location) version of the population $P_{org}$, as shown in Figure 2. Let $\mu_{QP_1}$ and $\mu_{QP_2}$ denote the means of the populations $P_{QP_1}$ and $P_{QP_2}$, respectively. Then, if these treatments have a systematic effect on video quality, we can express the elements of the corresponding populations as $P_{QP_1}^{(i)} = \mu_{org} + E_{QP_1} + \epsilon_i$ and $P_{QP_2}^{(i)} = \mu_{org} + E_{QP_2} + \epsilon_i$. Here $E_{QP_1}$ and $E_{QP_2}$ are the effects of the treatments $QP_1$ and $QP_2$, respectively. Hence, the quality scores for the new conditions are shifted from $\mu_{org}$ by an amount triggered by the visible impact of the treatments on the video quality, and can be quantified by $E_{QP_1}$ and $E_{QP_2}$. In the example shown in Figure 2, $E_{QP_1} = -1.2$ and $E_{QP_2} = -3.1$ (negative values are indicative of decrease in video quality). Notice that the resulting populations $P_{QP_1}$ and $P_{QP_2}$ will have the same variance as $P_{org}$ because the treatments ($QP_1$ and $QP_2$) will cause systematic changes in individual ratings (i.e. observers who were more critical in case of original video will remain so for the new conditions also). In the alternate case, if the treatments do not cause any changes in the opinion scores i.e the effect is not visible to the observers (i.e. $E_{QP_1} = 0$ and $E_{QP_2} = 0$), then the three populations will be the same and one can conclude that the treatments do not lead to statistically significant differences in means (MOS).

### B. The case of heterogeneous variances

In the third case, if the treatments $QP_1$ and $QP_2$ do not introduce systematic effect on video quality, then the individual opinion scores may randomly increase (video quality improves visibly according to some observers), decrease (video quality degrades visibly according to some observers) or remain the same (video quality levels remains same as without any treatment). In such case, we can say that the treatments caused the ratings to become heterogeneous because apart from the inherent random error ($\epsilon_i$), the varying values of $E_{QP_1}$ and $E_{QP_2}$ will introduce additional and possibly different variations in $P_{QP_1}$ and $P_{QP_2}$. Consequently, the variances of the three populations $P_{org}$, $P_{QP_1}$ and $P_{QP_2}$ will be different. Hence, testing if $\mu_{org} = \mu_{QP_1} = \mu_{QP_2}$ may not be useful since the populations will be different in any case. Practically, such cases are of less interest because one generally knows the effect of a given treatment *apriori* (in the given example of video compression, it is known $QP_1$ and $QP_2$ will lower video quality levels as compared to the original video) and statistical tests help to establish if the observed differences due to the treatment are merely due to chance (i.e. due to sampling error) or not.

If the population variances are unequal, it may point out to 2 possibilities: (1) additional factors may have crept in, (2) the observers have not been consistent in their ratings. The first possibility is generally minimized by careful experimental design including training sessions at the beginning of the test to ensure that the participants have understood the task well. The effect of second possibility is mitigated by rejecting outliers i.e. inconsistent observers that can cause variance to change are removed from further studies or analysis.

Such outlier rejection is well accepted and recommended in multimedia quality analysis, and well documented outlier rejection strategies exist [14], [20]. Therefore, outlier rejection provides indirect support for the assumption of homogeneity of variance, even though the explicit goal is to remove data points which might be *dissimilar* rather than making the variances of groups similar. In other words, experimental design in subjective tests for quality will help to ensure that the variances of the groups to be analyzed are similar. In general, the issue of heterogeneous group variances can be avoided [21] if proper experimental guidelines have been followed. In other words, Case 2 (i.e. samples/groups drawn from different populations with same population mean) may be practically less useful although it is perfectly valid for theoretical analysis. In summary, careful experimental design is more crucial for reliable statistical analysis and comparisons rather than focusing on homogeneity of variance and/or distributional assumptions (data normality).

It may also be noted that while the use of $t_{pooled}$, ANOVA requires that population variances are equal, it does not imply that sample/group variances be exactly equal. Rather the said variances should be similar. This can be quantified by computing the ratio of maximum to minimum group variance. Empirically, if the said ratio is greater than or less than $1/4\,(=0.25)$, then the population variances can be deemed to be unequal. In such case, it may not be meaningful to conduct $t$-test or ANOVA because the samples are likely to be drawn from two different populations.

### C. Comparing groups with different variances

Homogeneity of variance condition should be viewed in the light of practical considerations and not as a constraint. Therefore, it can be assessed via the empirical rule in order to obtain information about the presence of groups/samples that may have very different variances as compared to the remaining ones, and might suggest the possibility that the samples are taken from different populations (in which case comparing the means via $t_{unpooled}$ or other test which does not use pooled variance may be less meaningful). Once again, practical context should be used to ascertain if unequal variance condition is reasonable in view of the goals of analysis. For instance, it is possible that only a fraction of groups may violate this condition in which case the possible reasons can be examined. In other cases, such groups could possibly be removed from analysis. As discussed in section III-B, in theory $t_{pooled}$, ANOVA are in any case not affected by unequal variance if balanced design (equal sample size) is employed. Therefore experimental design should target balanced design as far as possible (in multimedia quality estimation, balanced design are common). Nevertheless, practically it may be more insightful to analyze the possible reasons and consequences of unequal variance rather than merely applying the statistical tests.

As discussed, Case 2 is valid from a theoretical perspective but is of less interest in practice. In other words, the implications of comparing $k$ samples whose corresponding populations have different variances but with equal means i.e. $\mu_1 = \mu_2 = ... = \mu_k$, should also be noted. In this context, it is useful to point out that MOS is sometimes not the most accurate measure of multimedia quality, and other measures may be required to supplement it. For instance, the authors in [22] proposed the use of SOS (standard deviation of opinion scores) while Ref. [23] suggested using PDU (percentage dissatisfied users) in addition to MOS. Note that measures such as SOS, PDU can be different even if corresponding population MOS are equal. Such cases will arise if groups (samples) from different populations (with same population means) are compared, and may not lead to meaningful analysis of perceptual quality and/or user satisfaction levels.

TABLE I: Description of distribution types and their characteristics.

| Type | Parameters | Shape | Kurtosis |
|---|---|---|---|
| Beta | $a = 0.5$,<br>$b = 0.5$ | symmetric,<br>bimodal (two peaks) | 1.5 |
| Exponential | $\lambda = 0.5$ | decaying curve,<br>non-symmetric | 9 |
| Normal | $\mu = 0$,<br>$\sigma = 1$ | bell-shaped, symmetric,<br>unimodal (one peak) | 3 |
| Uniform | $a = 0$,<br>$b = 1$ | flat (no peaks),<br>symmetric | 1.8 |

## V. Experimental results and discussion

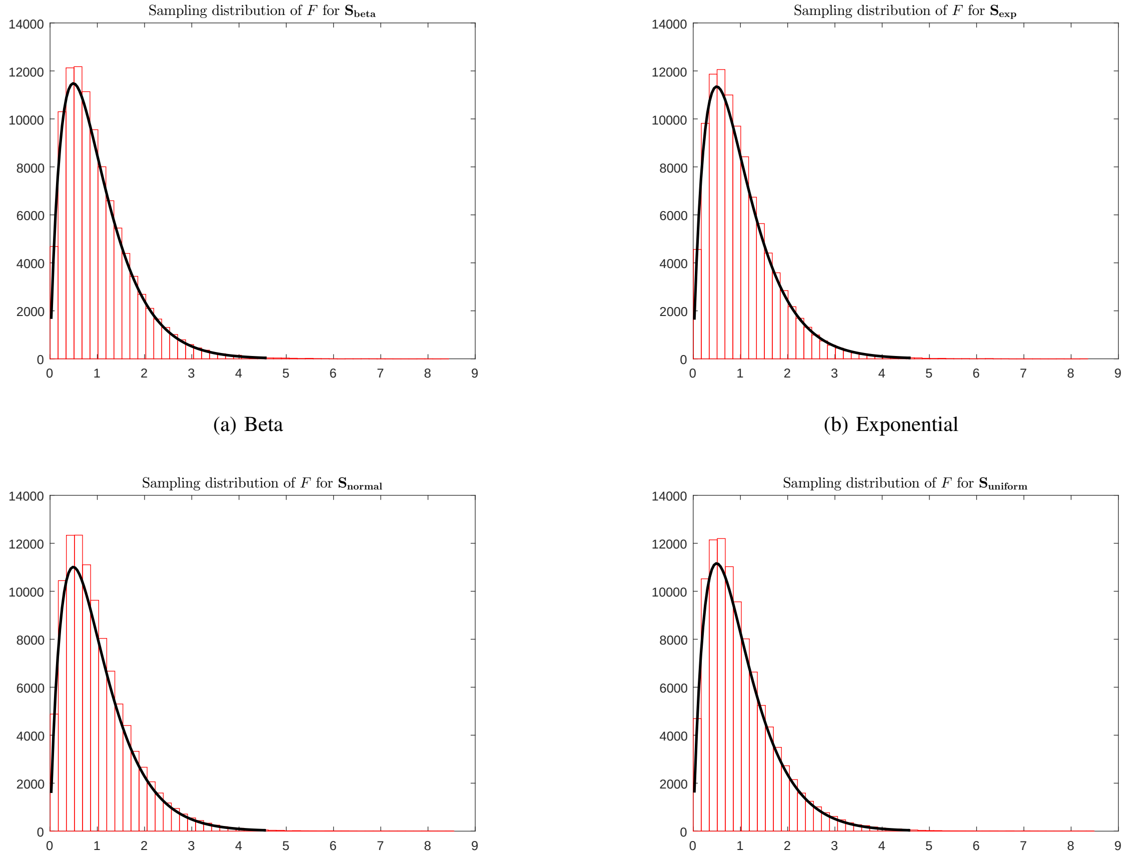
In the first set of experiments, we investigate the effect of type of distribution that the sample follows. We considered four different types of distributions (from which random numbers were generated to simulate sample observations), and these are summarized in Table I. Note that the parameters for these distributions were chosen in order to result in diverse shapes (in terms of symmetry, number of peaks etc.). The kurtosis values reported in Table I reflect this.

As an example, we use ANOVA, and study the sampling distribution of $F$ when the samples follow the distributions mentioned in Table I. We consider 5 groups ($k = 5$), equal number of observations in each group ($n_i = n = 25$), and ensured that the groups have similar variances. Thus, we represent the sample for exponential distribution as $\mathbf{S}_{exp} = [\mathbf{d}_{1exp}\ \mathbf{d}_{2exp}\ \mathbf{d}_{3exp}\ \mathbf{d}_{4exp}\ \mathbf{d}_{5exp}]$. Here $\mathbf{d}_{1exp}$ to $\mathbf{d}_{5exp}$ are 25-dimensional column vectors representing the groups. Similarly, we can define the samples for other distributions i.e. $\mathbf{S}_{beta}$, $\mathbf{S}_{normal}$ and $\mathbf{S}_{uniform}$.

Since our goal was to study the sampling distribution of $F$ in ANOVA, $\mathbf{S}_{exp}$, $\mathbf{S}_{beta}$, $\mathbf{S}_{normal}$ and $\mathbf{S}_{uniform}$ were generated randomly in each iteration, making sure the that observations followed the respective distributions. The sampling distributions of $F$ for each case are shown in Figure 3. The number of iterations $N_{iter} = 10^5$. We have also plotted (represented by continuous line) the theoretical $F$ distribution with the corresponding degrees of freedom i.e. $F(k-1, M-k) = F(4, 120)$ for comparison.

We can make the following two observations from this figure:

- The sampling distribution of $F$ follows the theoretical $F$-distribution curve irrespective of the type of sample distribution. Thus, sample normality is not a prerequisite for $F$ to be distributed according to $F$-distribution.



(a) Beta

(b) Exponential

(c) Normal

(d) Uniform

Fig. 3: Sampling distribution of $F$ values when the samples follow the indicated distributions. In each plot, the continuous curve indicates the theoretical $F$-distribution with 4 and 120 degrees of freedom. Figure best viewed in color.

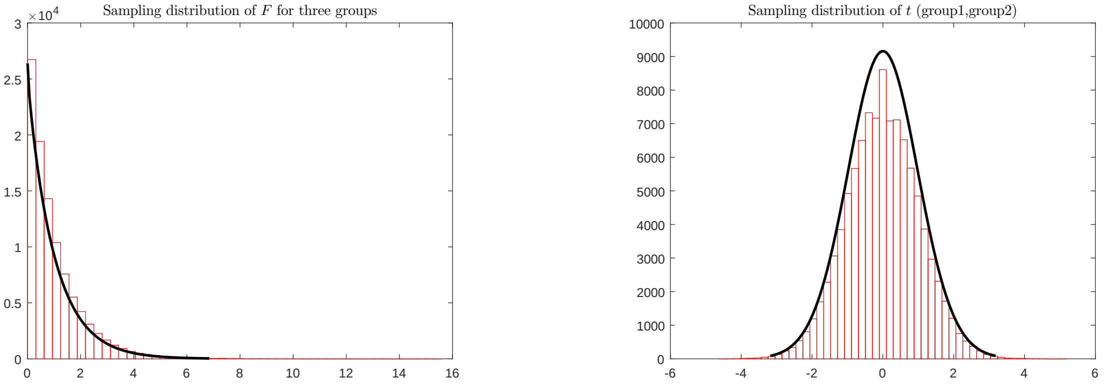
- Despite a small sample size ($n = 25$), the sampling distribution of $F$ approximates well the theoretical curve. Hence, as argued, in practice ANOVA (and other parametric tests) can be applied to approximate the theoretical distribution. Obviously, the approximations will improve with increasing sample size.

TABLE II: Description of groups taken from [5].

| | group 1 | group 2 | group 3 |
|---|---|---|---|
| Mean (MOS) | 5.5769 | 7.3846 | 7.3077 |
| Variance | 3.1338 | 3.2862 | 2.4615 |
| Kurtosis | 1.7971 | 6.9602 | 6.4978 |
| Shape | unimodal,<br>non-symmetric | bi-modal,<br>non-symmetric | unimodal,<br>non-symmetric |

We can carry out similar analysis regarding the sampling distribution of the test statistic on real data. However, in practice we typically have only one sample since the subjective or objective experiment is not repeated for obvious reasons. Therefore, to generate the sampling distributions in such scenario, we employ the idea of resampling. Specifically, given two or more samples which are to be compared, we can create randomized versions of these under the assumption that the given samples are similar (i.e. assuming the *null* hypothesis to be true). To demonstrate this, we use raw opinion scores from the dataset described in [5] where a comparison of upscalers was performed at varying compression rates. Since we want to study the sampling distribution of $F$ in ANOVA, we first selected three groups from the said data. These groups represent quality scores of three conditions evaluated by 26 observers. Thus, the group size was 26 ($n_i = n = 26$). Other descriptive properties of the selected groups are summarized in Table II from which we note that none of the groups are normally distributed as indicated by very high or very low kurtosis values and their shapes. In addition, the group variances are similar.

First, we applied ANOVA to compare the resampled versions of the three groups (we employed $10^5$ randomizations under the *null* hypothesis) and, the resulting sampling distribution of $F$ values is shown in Figure 4a. As expected, it approximates well the theoretical $F$ distribution. To give another example, we show the sampling distribution of $t_{pooled}$ when comparing group 1 and group 2 using the pooled $t$-test in Figure 4b. In this case also, the experimental distribution reasonably follows the theoretical $t$-distribution.



(a)

(b)

Fig. 4: Sampling distribution of $F$ and $t_{pooled}$ values for the groups of data taken from [5]. The groups are summarized in Table II. In each plot, the continuous curve indicates the corresponding theoretical distribution. Figure best viewed in color.

## VI. Practical recommendations

Based on the theoretical and experimental analysis in previous sections, it is clear that the application of parametric tests should focus on the consequences of the assumptions in these tests. The practical recommendations towards using the tests are highlighted in the right flow diagram in Figure 1, and are summarized in the following.

Applying normality checks on given data is neither required nor recommended as the CLT provides information about the shape and parameters of the sampling distribution of mean. Instead the more important consideration is whether mean (MOS) adequately represents the desired information from the sample(s). For instance, mean is a useful measure of central tendency in case of many symmetric distributions (not necessarily normal). Moreover, mean is still a practically useful statistic even if there are few outliers (skewness) in the data. In all such cases, parametric tests are practically meaningful for statistical analysis.

Homogeneity of variance should be exploited to obtain further insights into the data, and therefore not be viewed as a bottleneck for the purpose of statistical testing. To that end, the empirical rule (refer to section IV-B) should be applied to detect the presence of groups/samples that may have very different variances as compared to the remaining ones. If such groups exist, then the corresponding conditions should be revisited to find possible reasons for unequal variance. Consequently, if unequal variance condition is practically reasonable (or such groups can be removed), $t_{pooled}$ or ANOVA can be used. A balanced experimental design (equal sample size) would therefore be preferable in such cases (recall from section III-B both the tests are not affected by unequal variance if group/sample sizes are same).

The use of nonparametric tests is recommended if mean is not a suitable summary statistic of the data to be analyzed. Note that nonparametric tests should not be used merely because the given data is *nonnormal*. Rather they should be used to generate the sampling distribution of the desired test statistic.

In summary, analysis of data pertaining to multimedia quality using mean (average) as a test statistic should focus on experimental design (this includes the selection of challenging content recruiting adequate number of human subjects with possible emphasis on balanced design, conditions to be evaluated, and the final goal of analysis) rather than emphasizing distributional assumptions, equal variance condition or resorting to multiple hypothesis tests. However, if mean is not a suitable test statistic, then nonparametric tests can be used by leveraging the power of computers to construct empirical sampling distribution of the desired test statistic.

## VII. Concluding remarks

Parametric tests provide a theoretical framework for drawing statistical inferences from the data and thus help in formulating well grounded recommendations. However, the application of these tests and interpretation of the results require some care in the light of the assumptions required in these tests. To that end, we revisited the theoretical formulations and clarified the role of the assumption of normality and homogeneity of variance. By analyzing the sampling distribution of the test statistics, we argued that the more appropriate question to be asked before deploying parametric tests is whether the test statistic follows the corresponding distribution or not (instead of the data following any specific distribution). We also emphasized that the said assumptions should not be viewed as constraints on the data. Instead it is more important to focus on their practical implications.

The presented analysis is particularly relevant in the context of multimedia quality assessment because the said issues have not been emphasized enough in the corresponding literature. We also made practical recommendations in order to avoid the theoretical issues related to multiple hypothesis testing. Even though the targeted application was multimedia quality estimation, the theoretical arguments and the recommendations are expected to be useful in several other areas (such as medical data analysis, information retrieval, natural language processing etc.) where parametric tests are widely used. In order to provide a tool for practical use, a software implementing the said recommendations is also made publicly available[§].

## References

[1] P. Coverdale, S. Moller, A. Raake, and A. Takahashi, “Multimedia quality assessment standards in itu-t sg12,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 91–97, Nov 2011.

[2] ITU-R Recommendation BS.1534-3, “Method for the subjective assessment of intermediate quality levels of coding systems,” International Telecommunication Union, Geneva, Switzerland, Tech. Rep., Oct. 2015.

[3] ITU-T Tutorial, “Objective perceptual assessment of video quality: Full reference television,” International Telecommunication Union, Geneva, Switzerland, Tech. Rep., May 2005.

[4] T. K. Tan, R. Weerakkody, M. Mrak, N. Ramzan, V. Baroncini, J. R. Ohm, and G. J. Sullivan, “Video quality evaluation methodology and verification testing of hevc compression performance,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 76–90, Jan 2016.

[5] Y. Pitrey, M. Barkowsky, P. Le Callet, and R. Pepion, “Subjective quality evaluation of h.264 high-definition video coding versus spatial up-scaling and interlacing,” *ACM EuroITV Conference, Workshop on Quality of Experience for Multimedia Content Sharing (QoEMCS)*, 2010.

[6] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, “Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality,” *Optical Engineering*, vol. 52, no. 10, pp. 102 008–102 008, 2013.

[7] B. Belmudez, *Audiovisual Quality Assessment and Prediction for Videotelephony*, ser. T-Labs Series in Telecommunication Services. Springer International Publishing, 2016.

[8] M. Melo, M. Bessa, L. Barbosa, K. Debattista, and A. Chalmers, “Screen reflections impact on hdr video tone mapping for mobile devices: an evaluation study,” *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, p. 44, 2015.

[9] J. M. Moreno-Roldán, M. Luque-Nieto, J. Poncela, V. Díaz-del-Río, and P. Otero, “Subjective quality assessment of underwater video for scientific applications,” *Sensors*, vol. 15, no. 12, pp. 31 723–31 737, 2015. [Online]. Available: http://dx.doi.org/10.3390/s151229882

[10] F. M. Moss, K. Wang, F. Zhang, R. Baddeley, and D. R. Bull, “On the optimal presentation duration for subjective video quality assessment,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 1977–1987, Nov 2016.

[11] ITU-T Recommendation P.1401, “Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models,” International Telecommunication Union, Geneva, Switzerland, Tech. Rep., Jul. 2012.

[12] G. Pflug, “On kersting’s proof of the central limit theorem,” *Statistics & Probability Letters*, vol. 1, no. 6, pp. 323 – 326, 1983.

[13] G. Roussas, *An Introduction to Probability and Statistical Inference*, second edition ed. Academic Press, 2015.

[14] ITU-R Recommendation BT.500-12, “Methodology for the subjective assessment of the quality of television pictures.” Geneva, Switzerland: International Telecommunication Union, 2009.

[15] M. H. Pinson, L. Janowski, R. Pepion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. L. Callet, M. Barkowsky, and W. Ingram, “The influence of subjects and environment on audiovisual subjective tests: An international study,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 640–651, Oct 2012.

[16] E. Schmider, M. Ziegler, E. Danay, L. Beyer, and M. Bühner, “Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption.” *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, vol. 6, no. 4, p. 147, 2010.

[17] W. K. Lim and A. W. Lim, “A comparison of usual t-test statistic and modified t-test statistics on skewed distribution functions,” *Journal of Modern Applied Statistical Methods*, vol. 15, no. 2, pp. 67–89, 2016.

[18] T. Lumley, P. Diehr, S. Emerson, and L. Chen, “The importance of the normality assumption in large public health data sets,” *Annual Review of Public Health*, vol. 23, pp. 151–169, 2002.

[19] B. L. Welch, “The significance of the difference between two means when the population variances are unequal,” *Biometrika*, vol. 29, no. 3-4, p. 350, 1938.

[20] “Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment.” Video Quality Experts Group (VQEG), March 2003.

[21] S. S. Sawilowsky, “Fermat, schubert, einstein, and behrens-fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$,” *Journal of Modern Applied Statistical Methods*, vol. 1, no. 2, pp. 461–472, 2002.

[22] T. Hofeld, R. Schatz, and S. Egger, “Sos: The mos is not enough!” in *2011 Third International Workshop on Quality of Multimedia Experience*, Sept 2011, pp. 131–136.

[23] D. C. Mocanu, J. Pokhrel, J. P. Garella, J. Seppnen, E. Liotou, and M. Narwaria, “No-reference video quality measurement: added value of machine learning,” *Journal of Electronic Imaging*, vol. 24, no. 6, p. 061208, 2015. [Online]. Available: http://dx.doi.org/10.1117/1.JEI.24.6.061208

[§] https://sites.google.com/site/narwariam/home/research