# Evaluation of Statistical Inference Tests Applied to Subjective Audio Quality Data With Small Sample Size

Jeroen Breebaart

***Abstract*—Monte-Carlo simulations of statistical inference tests were performed to assess type 1 (false rejection) and type 2 (false non-rejection) error rates associated with subjective audio quality data as a function of sample size. Samples were generated by randomly drawing data from large-scale subjective audio quality tests. Null hypotheses were simulated by equalizing population means followed by pooling. The Null hypothesis rejection rates were determined for a parametric $t$ test, as well as a non-parametric (permutation) test and compared to rejection rates based on analytical expressions and empirical distributions of the sample means and medians. The results indicated that pairwise comparisons are beneficial for high power and to obtain type I error rates that are close to the nominal value of 5%. The pairwise inferences can be realized by a parametric, pairwise $t$ test or by a non-parametric permutation test, provided that for the latter, only pairwise permutations are executed. Although the observations from this study cannot be generalized for arbitrary data sets, the results do indicate that a pairwise, non-parametric resampling test is an interesting candidate for the statistical analysis of subjective quality data due to the absence of any requirements on data distributions and its relatively accurate performance in terms of Null hypothesis rejection rates.**

***Index Terms*—Audio coding, audio compression, Monte Carlo methods, statistical analysis.**

## I. Introduction

THE development, verification and standardization of audio codecs has in many cases been supported by double-blind, subjective listening tests such as defined in ITU-R BS.1534 [1], [2]. In such tests, assessors rate the perceived quality of one or more codecs against a reference. The resulting quality ratings, expressed in MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) points, may sometimes raise doubts whether one system is outperforming another, due to dependencies on the audio excerpt under test, the assessor familiarity with the test procedure, and general variability in subjective responses. ITU-R BS.1534-1 [1] therefore recommends the use of statistical intervals that indicate the 95% confidence region for the mean of each system under test based on the Student's $t$ distribution. Furthermore, the document recommends to provide "information about the inherently statistical nature of all subjective data", but does not describe how such information should be derived. A recent revision of ITU-R BS.1534-2 [2] provides more elaborate instructions on reporting statistical structures by means of parametric inferences, and additionally includes handles to non-parametric inference tests. Generally speaking, statistical inferences estimate the probability of finding extreme differences given a postulated Null hypothesis. In traditional, parametric methods, this Null hypothesis is often formulated as two or more systems presumably having equal population means, and evidence to reject this Null hypothesis is typically based on assumed parametric probability density functions of the sample mean such as the binomial distribution [3] or the Student's $t$ distribution [4].

Parametric statistical inference tests come with certain requirements that the data should adhere to. For example, the Student's $t$ test is ideally designed for independent samples that follow a Gaussian distribution and have equal variance in groups that are being compared. Violation of such requirements can give inaccurate results, although the $t$ test has been shown to be reasonably robust against violations in normality [5]–[8]. Alternatively, non-parametric tests can be employed such as the (random) permutation test [2], [9], [10]. The permutation test is not based on the sampling distribution of the mean, nor does it impose any requirements on the distribution of the data. The Null hypothesis associated with a permutation test is an exchangeability requirement, that is, the coupling of the dependent variable (e.g. the subjective quality score) with the independent variable(s) (codec, item and/or assessor under test) is random [11]. In this respect, the Null hypothesis is stronger than the assumption of equal means associated with the $t$ test. The consequence is that the permutation test also tends to respond to changes in the distributions other than a difference in the mean [12]–[15].

The purpose of this paper is to report on an attempt to evaluate the accuracy of the most commonly-used parametric statistical inference test (Student's $t$) and a non-parametric test (permutation test) when applied to small sets of subjective audio quality data. The study is based on the assumption that large populations of measurement data are known a priori. For this purpose, large scale, publicly-available listening tests results published by the Moving Picture Experts Group (MPEG) are assumed to represent "true" populations of subjective quality assessments for a set of audio codecs. Small subsets of the population were randomly sampled to simulate the practical process of random testing with a limited number of subjects. These small samples were subjected to a variety of statistical inference tests. The accuracy of these inference tests was evaluated in terms

Manuscript received October 09, 2014; revised February 02, 2015; accepted March 06, 2015. Date of publication March 11, 2015; date of current version March 27, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wai-Yip Geoffrey Chan.

The author resides in Pyrmont, NSW 2009, Australia (e-mail: jeroen@jeroenbreebaart.com).

Digital Object Identifier 10.1109/TASLP.2015.2412462

of their ability to (in)correctly reject the Null hypothesis. For this purpose, the rejection rates were determined by means of Monte-Carlo simulations and compared to rejection rates based on (1) the test's confidence level (or $p$ value), (2) analytical expressions of the sample mean (or median), and (3) the empirically-derived distribution of the sample mean (or median). The methods to calculate the rejection rates are outlined in Section II, and the comparisons of results are given in Section III. A discussion and conclusions are provided in Sections IV and V, respectively.

## II. Method

### A. Analytical derivation of power and type 1 error (false reject) probability

Throughout this paper, we are interested in comparing two sets of samples, $\{y_1, \dots y_N\}$ and $\{z_1, \dots z_N\}$, both with cardinality $N$. These sets of samples are drawn from populations with known, real-valued, continuous ($Y, Z \in \mathbf{R}$), one-dimensional probability density functions (PDFs), denoted by $f_Y(y)$ and $f_Z(z)$, respectively. A set of $N$ randomly drawn samples from a population with probability density function $f_Y(y)$ has a sum $Y_N$. The PDF of that sum, $f_{Y_N}(y)$, assuming independent samples $y_n$, is obtained by repeated convolution [16]:

$$f_{Y_N}(y) = f_Y^{\otimes N}(y), \tag{1}$$

with $(.)^{\otimes N}$ the repeated convolution operator:

$$f^{\otimes N}(y) = \left( \underbrace{f \otimes \dots \otimes f}_{N} \right)(y), \tag{2}$$

and

$$(f \otimes g)(y) = \int_{-\infty}^{\infty} f(a) g(y-a) da. \tag{3}$$

By substitution of $y' = y/N$ it then follows that the probability density function $f_{\bar{Y}_N}(y')$ of the arithmetic sample mean $\bar{Y}_N$ is given by [16]

$$f_{\bar{Y}_N}(y') = \left| \frac{\partial y}{\partial y'} \right| f_{Y_N}(y) = N f_{Y_N}(Ny'). \tag{4}$$

If the location statistic of interest is the median across $N$ samples, the PDF $f_{\tilde{Y}_N}(y)$ for the median $\tilde{Y}_N$ for odd $N$ is given by the multinomial probability function of $N$ trials resulting in $m$ samples below $y$, $m$ samples above $y$, and one sample at $y$ with $m = (N-1)/2$:

$$f_{\tilde{Y}_N}(y) = \binom{N}{m, m, 1} (\Pr(Y < y))^m \times (\Pr(Y > y))^m \Pr(Y = y), \tag{5}$$

which is equal to

$$f_{\tilde{Y}_N}(y) = \frac{N!}{(m!)^2} (c_Y(y))^m (1 - c_Y(y))^m f_Y(y), \tag{6}$$

with $c_Y(y)$ the cumulative probability density function of $f_Y(y)$:

$$c_Y(y) = \int_{-\infty}^{y} f_Y(a) da. \tag{7}$$

Because the sample location statistic (e.g., the sample mean or median) is a stochastic variable, the difference between two location statistics is a stochastic variable as well. In particular, when comparing two sample means $\bar{Z}_N$ and $\bar{Y}_N$:

$$\bar{X}_N = \bar{Z}_N - \bar{Y}_N, \tag{8}$$

the probability density function of $\bar{X}_N$ is obtained by correlation

$$f_{\bar{X}_N}(x) = \left( f_{\bar{Y}_N} \star f_{\bar{Z}_N} \right)(x), \tag{9}$$

which is equivalent to the convolution of $f_{\bar{Z}_N}(a)$ with $f_{\bar{Y}_N}(-a)$:

$$f_{\bar{X}_N}(x) = \int_{-\infty}^{\infty} f_{\bar{Z}_N}(a) f_{\bar{Y}_N}(a - x) da. \tag{10}$$

The same expression can be used for the difference between two sample medians $\tilde{Z}_N$ and $\tilde{Y}_N$.

Alternatively, one may be interested in the distribution of pairwise differences, and in this particular case, the PDF of pairwise differences $X$ can be derived from the population of pairwise differences $x_i = z_i - y_i$. Subsequently, the PDF of the mean of $N$ samples $x_i$ can be derived from $f_X(x)$ based on Eqs. (1)–(4).

Irrespective of whether pairwise or sample-wise differences are calculated, a Null hypothesis $H_0$ can be postulated that assumes that the samples $y_i$ and $z_i$ are randomly drawn from the same underlying population. In the context of a subjective quality evaluation, this implies that two different codecs result in the same subjective quality and that differences in subjective scores between codecs are to be treated as measurement noise. The alternative hypothesis $H_1$ assumes that the PDFs of $Y$ and $Z$ are not equal, for example because their means, variances, or shapes of their PDFs differ.

Traditionally, statistical inference tests provide handles to two types of errors given a set of samples. The first error, often referred to as a type 1 error, is the error of rejecting the Null hypothesis while in fact that hypothesis is true. In other words, the type 1 error is a false rejection (of the Null hypothesis) and we will use that term in conjunction with the use of type 1 for clarity. The probability that this error occurs is denoted $\alpha$, and in practice it the probability that a certain range of (extreme) data are observed under the Null hypothesis. For a one-sided test (i.e., if we are only interested in differences in a certain direction, for example that the true population mean is larger than zero), $\alpha$ can be expressed as a function of a critical value $x_t$:

$$\alpha_{\bar{X}_N|H_0}(x_t) = 1 - c_{\bar{X}_N|H_0}(x_t). \tag{11}$$

If the sample mean is found to be larger than or equal to a predetermined $x_t$ corresponding with a desirable level of significance $\alpha$, the data are considered as evidence that the Null hypothesis is false and that hypothesis is therefore said to be rejected. As $\alpha$ is a *conditional* probability under $H_0$, it cannot be interpreted as a probability that the Null hypothesis is true. Instead, if the sample mean is found not to exceed $x_t$, the test is interpreted as being *inconclusive*. In a similar way, a type 2 (false non-rejection) error, or $\beta$ can be computed as the probability of finding a sample location statistic smaller than the critical value $x_t$ under $H_1$:

$$\beta_{\bar{X}_N|H_1}(x_t) = c_{\bar{X}_N|H_1}(x_t). \tag{12}$$

$\beta$ represents the probability of not rejecting the Null hypothesis associated with critical value $x_t$ if the alternative hypothesis $H_1$ is true, and is often referred to as the false negative, or false non-reject probability. Due to their inverse dependency on $x_t$, $\alpha$ and $\beta$ are subject to a trade-off; a larger value of $x_t$ results in a smaller $\alpha$ at the cost of an increasing $\beta$. Furthermore, an increase in the sample size $N$ reduces $\alpha$, $\beta$, or both. Last but not least, the sample mean and median have different distributions, and therefore the critical value $x_t$ to reject the Null hypothesis is different for these two metrics of central tendency. For example, for a Gaussian distribution, the variance of the sample median is approximately 1.6 times larger than the variance of the sample mean [17].

Common practice is to estimate the type 1 (false rejection) error probability given observed data, or more precisely based on a location statistic derived thereof (cf. [18], [19]). The result of this method is commonly denoted by the $p$ value, which is an estimate of the probability of finding data at least as extreme as observed in the test if the Null hypothesis is true. As indicated in the previous sections, for a finite sample size $N$, sample location statistics are stochastic variables, and consequently, the $p$ value is a stochastic variable as well. If (estimates of) the probability density functions of $\bar{X}_N$ under $H_0$ and $H_1$ are known, one can analytically compute the probability density function for the $p$ value. Under $H_1$, the PDF for the mean (or median) across $N$ samples is denoted $f_{\bar{X}_N|H_1}(x)$. The $p$ value corresponding to that mean, e.g., for a one-sided test, is given by a function $g(x)$ consisting of the integral of $f_{\bar{X}_N|H_0}(x)$ evaluated under $H_0$:

$$p = g(x) = \int_{a=x}^{\infty} f_{\bar{X}_N|H_0}(a)da = 1 - c_{\bar{X}_N|H_0}(x). \tag{13}$$

Given the transform function $p = g(x)$ of Eq. (13) and assuming $g(x)$ is a monotonously increasing or decreasing function with $g(x) > 0$, we can express the PDF of $p$, $f_P(p)$, by applying a PDF transformation (cf. [16, p. 200])

$$f_P(a) = \left| \frac{\partial g^{-1}(a)}{\partial p} \right| f_{\bar{X}_N|H_1}(g^{-1}(a)). \tag{14}$$

Applying the chain rule for differentiating inverse functions gives

$$f_P(a) = \left| \frac{\partial g(x(a))}{\partial x} \right|^{-1} f_{\bar{X}_N|H_1}(x(a)), \tag{15}$$

which results in

$$f_p(a) = \frac{f_{\bar{X}_N|H_1}(x(a))}{f_{\bar{X}_N|H_0}(x(a))}. \tag{16}$$

This formulation implies that if the populations under $H_0$ and $H_1$ have equal probability density functions and assuming independently-drawn data, all values of $p$ are equally probable (e.g., $f_p(a) = 1$ for $0 \leq a \leq 1$). If the PDFs of $\bar{X}_N$ under $H_0$ and $H_1$ are not equal, on the other hand, the power of a test is equal to the probability of finding a location statistic at or above the critical value $x_t(\alpha)$ under $H_1$:

$$\text{Power}(\alpha) = 1 - \beta(x_t(\alpha)) = 1 - c_{\bar{X}_N|H_1}(x_t(\alpha)). \tag{17}$$

Throughout the remainder of the paper, $H_0$ rejection rates computed as described in this section will be referred to as the method using an *analytical* PDF, or "APDF" in short.

### B. Empirical Derivation of Power and Type 1 (false rejection) Error Probability

The analytical expressions given in the previous section to compute power as a function of the sample size $N$ for arbitrary population distributions assumes that samples are randomly drawn and independent. This may not always be true, and therefore it is of interest to compare the analytical results against power and type 1 (false rejection) errors for empirically-determined distributions of the mean or median. These distributions can be obtained by Monte-Carlo sampling under $H_0$ and $H_1$. Assuming $K$ Monte-Carlo iterations, a one-sided critical value $x_t$ given a confidence level $\alpha$ is then obtained by satisfying

$$\alpha = \frac{1}{K} \sum_{k=1}^{K} I\left(\hat{X}_{N,k|H_0} \geq x_t\right), \tag{18}$$

with $I(.)$ the indicator function, and $\beta$ subsequently given by

$$\beta = \frac{1}{K} \sum_{k=1}^{K} I\left(\hat{X}_{N,k|H_1} < x_t\right), \tag{19}$$

with $\hat{X}_{N,k|H_i}$ the sample mean across $N$ samples for Monte-Carlo iteration $k$ under hypothesis $H_i$. The same procedure can be used for the sample median. Throughout the remainder of this paper, $H_0$ rejection rates computed using the method described in this section will be referred to as using an *empirical* PDF, or "EPDF" in short.

### C. Parametric estimation of power and type 1 (false rejection) error

In practice, the population distributions of $Y$ and $Z$ are unknown, and therefore only *estimates* of $p$ can be computed, which, besides the inevitable variability in $p$ due to its stochastic nature resulting from finite $N$, may also comprise estimation errors as a result of assumptions made in the estimation process. For example, the well-known independent (or two-sample) Student's $t$-test is based on data with a presumed Gaussian distribution, independent samples and equal variance. A modification to partly compensate for unequal variance is Welch's $t$ test. Based on a Null hypothesis postulating equal population means (e.g., $\mu_Z = \mu_Y$) the statistic $T_w$ for two sets of $N$ samples $\{y_1, \ldots y_N\}$ and $\{z_1, \ldots z_N\}$ is computed according to:

$$T_s = \frac{\bar{Z}_N - \bar{Y}_N}{\sqrt{N^{-1}\left(s^2_{Z_N} + s^2_{Y_N}\right)}}, \tag{20}$$

with $s^2_{Z_N}$, $s^2_{Y_N}$ the unbiased estimators of the variances of their respective populations. The resulting $p$ value is computed as the area under the cumulative $t$ distribution with $\nu$ degrees of freedom derived using the Welch-Satterthwaite equation [20] for equal sample sizes and unequal variances:

$$\nu = (N-1)\,\frac{\left(s_{Y_N}^2 + s_{Z_N}^2\right)^2}{s_{Y_N}^4 + s_{Z_N}^4}. \tag{21}$$

If samples $y_n, z_n$ are expected to have a certain degree of correlation, as can be the case in subjective audio quality tests due to a dependency on assessor sensitivity and/or differences in how critical audio test items are, such common factors can be partly canceled out by applying a pairwise $t$ test. In this test, pairwise differences are calculated of the two sets of $N$ samples $\{y_1, \dots y_N\}$ and $\{z_1, \dots z_N\}$ with a Null hypothesis that the expected difference is zero:

$$x_n = z_n - y_n, \tag{22}$$

and the resulting test statistic $T_p$ is based on $N-1$ degrees of freedom and calculated according to:

$$T_p = \frac{\bar{X}_N \sqrt{N}}{s_{X_N}}. \tag{23}$$

The type 1 (false rejection) and type 2 (false non-rejection) error rates can be established by computing $p$ values associated with $T_s$ or $T_p$ across $K$ Monte-Carlo iterations with the $t$ test operating at a nominal $\alpha$ value set to 0.05, and determining the Null-hypothesis rejection rate for data sampled under $H_0$ and $H_1$, respectively.

### *D. Non-Parametric Estimation of Power and Type 1 (false rejection) Error Probability*

An interesting class of non-parametric statistical inference tests is the permutation test. In this test, the Null hypothesis postulates exchangeability of group labels and the associated $p$ value is computed by permuting data between the two sets $\{y_1, \dots y_N\}$ and $\{z_1, \dots z_N\}$ and calculating the proportion of those permutations resulting in some differential sample statistic equal or larger than the difference in that location statistic obtained from the two sets without permutation. Under the assumption that each permutation is equally probable, that proportion is equal to the conditional probability under $H_0$ of finding an effect as large, or larger than the one found prior to permutation. The permutations can be applied across all data available in the two sets, and we denote this full permutation configuration by $P_f$. Alternatively, one can decide to maintain assessor and item dependencies in the permutation test by only applying the pairwise permutation subset, e.g., by only (randomly) permuting corresponding pairs $z_n, y_n$. This pairwise permutation method is denoted by $P_p$. Both the mean as well as the median can be employed as sample location statistic. One can compute permutation test $p$ values for randomly-drawn samples from a population, and investigate the power of the test under $H_1$ and the type 1 (false rejection) error rate under $H_0$ to allow comparison with the conventional, parametric tests in a similar approach as outlined in [21], [22].

The $p$ value in a single permutation test is lower bounded by the reciprocal of the number of permutations $Q$ (e.g. $5 \times 10^{-5}$ for $Q = 20000$). The relative error of that $p$ value, assuming a binomial distribution and a sufficient number of independent observations, can be estimated using

$$\frac{\sqrt{Var(p)}}{\bar{p}} \approx \frac{\sqrt{\bar{p}(1-\bar{p})/Q}}{\bar{p}} \approx \frac{1}{\sqrt{\bar{p}Q}}. \tag{24}$$

This means that for a $p$ value of 0.05, and $Q = 20000$ permutations, the standard error of $p$ is about 31.6 times smaller than the $p$ value itself.

### *E. Monte-Carlo Procedure*

Monte-Carlo sampling from datasets comprised random selection of a combination of test item and assessor, and subsequently assigning the associated assessor responses to $y_n$ and $z_n$, for the two systems that are being compared. For every condition, $K = 10000$ Monte-Carlo iterations were employed. The sample size $N$ was either 3, 5, 7, 9, 11, 15, 19, 23 or 31. Permutation tests were employed with $Q = 20000$ permutations, unless the unique number of permutations was smaller; in that case, all unique combinations were tested once.

### *F. Null Hypothesis*

In all simulations, the Null hypothesis postulates that there is no consistent difference between the scores of the two codecs. Data under the Null hypothesis were created by linearly translating the data of $Z$ to equalize its population mean to that of $Y$, followed by pooling of the data to ensure that the data PDFs (and hence any location statistic derived thereof) of $Y$ and $Z$ were identical under $H_0$ to meet the exchangeability requirement. In all configurations, a nominal value $\alpha$ of 0.05 was used [2]. Only one-sided tests were evaluated.

## III. Results

### *A. MPEG Surround Verification Test*

*1) Data Sets:* In January 2007, the MPEG audio group published the results of an MPEG Surround verification test [23] demonstrating the performance of the MPEG Surround multi-channel codec [24], [25] at various bit rates. The data sets used from this test are summarized in Table I. The first and second column indicate the acronym used in this paper, and the full condition name used in [23], respectively. The test procedure according to ITU-R BS.1534-1 [1] was employed with 10 audio excerpts and 41 listeners distributed across 4 laboratories, resulting in 410 responses for each codec configuration. The subset of codecs used in the current study are given in Table I. The first column in the table denotes the configuration acronym used throughout this paper; the second column describes the full test configuration as used in [23]. The mean, standard deviation, skew and excess kurtosis for the subjective (MUSHRA) scores associated with each codec are given in the various columns of Table I. Subjective scores do not only differ in their means across codecs; additionally variation in their standard deviations exists as well ranging from 2.0 for REF to 19.8 MUSHRA points for DPL2. Last but not least, REF and MPS-A show skew and excess kurtosis substantially different from zero, indicating that the distribution of these data deviates from a Gaussian distribution.

TABLE I
SUBJECTIVE (MUSHRA) SCORE SAMPLE SIZE, MEAN, STANDARD DEVIATION, SKEW AND KURTOSIS (PEAKEDNESS) FOR EACH CODEC SELECTED FROM THE MPEG SURROUND VERIFICATION TEST [23]. THE COLUMN "ACRONYM" INDICATES THE ACRONYM USED THROUGHOUT THIS PAPER; THE COLUMN "MPEG TEST/SYSTEM" CONTAINS THE FULL TEST AND CONDITION NAME USED IN [23]. KURTOSIS IS EXPRESSED AS EXCESS KURTOSIS (E.G., THE KURTOSIS CALCULATED FROM THE DISTRIBUTION OF THE DATA MINUS THE KURTOSIS OF A GAUSSIAN DISTRIBUTION)

| Acronym | MPEG test/system | Size | Mean | Std | Skew | Excess kurtosis |
|---|---|---|---|---|---|---|
| REF | MPEG Surround T1mc REF | 410 | 99.4 | 2.0 | -4.8 | 29.8 |
| MPS-A | MPEG Surround T1mc HE-AAC MPS A | 410 | 90.4 | 10.9 | -1.8 | 3.8 |
| MPS-B | MPEG Surround T1mc HE-AAC MPS B | 410 | 75.8 | 18.6 | -0.8 | 0.1 |
| DPL2 | MPEG Surround T1mc L2 DPL2 | 410 | 56.9 | 19.8 | +0.2 | -0.7 |

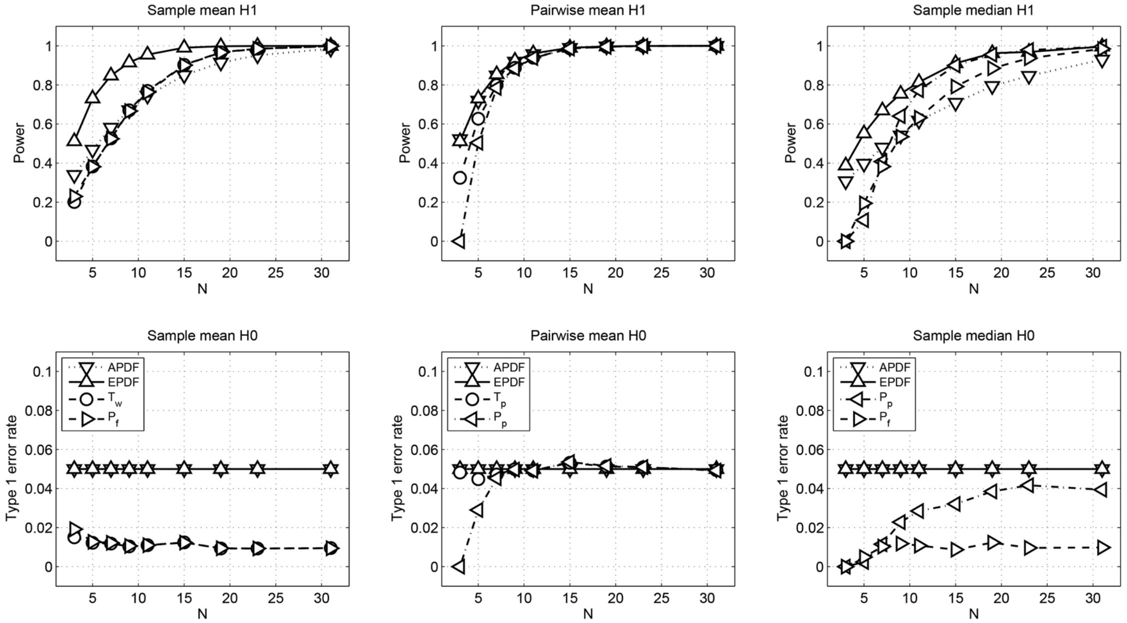

Fig. 1. Comparison of $H_0$ rejection rates for DPL2 vs MPS-B. The upper panels and lower panels evaluate power under $H_1$ and type 1 (false rejection) error rates under $H_0$, respectively. From left to right, the three panels represent the sample mean, pairwise mean, and sample median as location statistic. Different symbols denote the various inference methods; downward triangles ("APDF") refer to the analytical PDFs of the location statistic and upward triangles ("EPDF") refer to empirically-determined PDFs. Circles denote results for the $t$ test on sample means ("$T_w$") or pairwise means ("$T_p$"). Leftward and rightward triangles refer to the permutation test using sample means and pairwise differences, respectively.

*2) DPL2 vs MPS-B:* When comparing conditions DPL2 and MPS-B, the scores for the two codecs under test have very similar standard deviations and reasonably small skew (see Table I). The results of the Monte-Carlo simulations are shown in Fig. 1. The top panels correspond to simulations under $H_1$ showing power as a function of the sample size $N$. The bottom panels show the type 1 (false rejection) error rates under $H_0$ as a function of $N$. Different symbols represent the various inference methods (see legends). From left to right, the three panels represent the use of the sample mean, pairwise mean, and sample median, respectively. When the results under $H_1$ are considered (upper panels), it can be observed that, expectedly, power increases with $N$. For all methods, $N = 31$ results in a power equal to, or very close to $+1$, but the convergence differs considerably between methods. For tests on the sample mean (top-left panel), the empirical distribution of the mean (upward triangles) provides the highest power. The analytical distribution of the sample mean (downward triangles), the parametric $t$ test (circles) and the permutation test (rightward triangles) perform very similar but provide lower power than the empirical distribution. For pairwise comparisons using the mean (top-middle panel), all methods provide a very similar power as a function of $N$ except for some differences with very small sample sizes ($N \leq 7$), for which the power of the permutation test is lower bounded due to the limited number of unique permutations. Furthermore, the pairwise comparisons provide higher power than their counterparts based on a comparison of sample means, with the exception of the empirical distribution of the mean. When the median as location statistic is considered (right top panel), the power associated with pairwise permutations (leftward triangles) is higher than that for the full set of permutations (rightward triangles), and for $N \geq 11$ approaches the power of the empirical distribution (upward triangles). Nevertheless, the power associated with the median as location statistic is typically lower than the power for the corresponding tests using the mean, especially with pairwise tests.

The results under $H_0$ (lower panels) indicate that the type 1 (false rejection) error rates (e.g., the proportion of Monte-Carlo iterations under $H_0$ in which the Null hypothesis was erroneously rejected) are typically smaller than the nominal value of 0.05, except for the pairwise comparisons with the mean as location statistic (middle panel). Tests employing the sample mean (bottom-left panel) are quite conservative with $H_0$ rejection rates between 0.01 and 0.02, and are similar to those observed for the permutation test with the sample median (rightward triangles in the bottom-right panel). The type 1 (false rejection) error rates for pairwise permutations with the sample median (leftward triangles in the bottom-right panel)

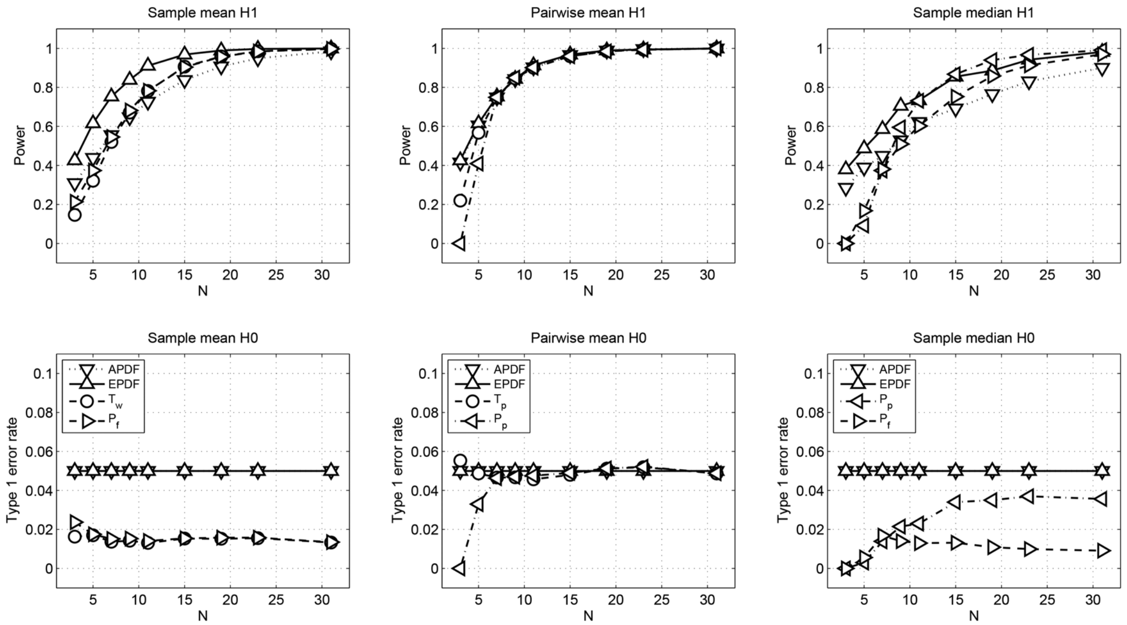

Fig. 2. Comparison of MPS-B and MPS-A. Same layout as in Fig. 1.

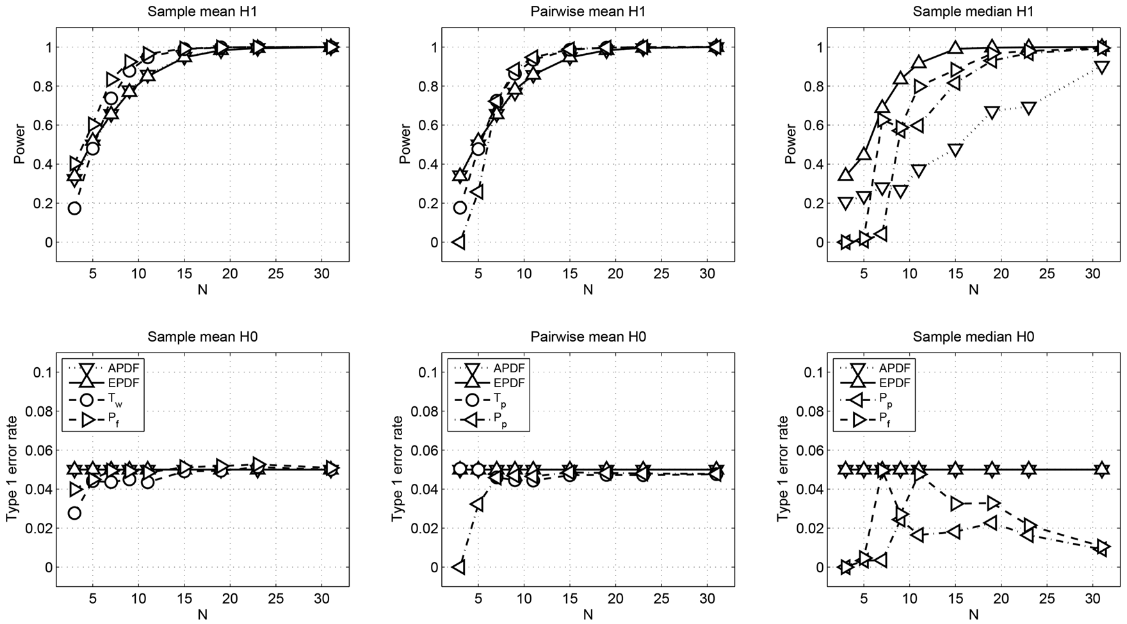

Fig. 3. Comparison of MPS-A vs REF. Same layout as in Fig. 1.

are closer to the nominal $\alpha = 0.05$ for $N \geq 9$, converging to about 0.04 with increasing $N$.

*3) MPS-B vs MPS-A:* The standard deviations of the populations of MPS-B and MPS-A differ by almost a factor of two (cf. Table I) and therefore this comparison is of interest to study the effect of nonuniform variance among two data sets. The results obtained for the comparison between MPS-B and MPS-A are shown in Fig. 2. Except for minor quantitative differences, the results are qualitatively quite similar to those observed in Fig. 1.

*4) MPS-A vs REF:* The comparison of MPS-A against the hidden reference REF is of particular interest, since the standard deviations of the two distributions differ considerably, and additionally, the scores for the hidden reference show considerable skew and kurtosis. These properties therefore violate data assumptions associated with the $t$ test. The power and type 1 (false rejection) error rates as a function of the sample size $N$ are shown in Fig. 3. When power is considered (upper panels), the analytical (downward triangles) and empirical (upward triangles) distributions show virtually identical results for the mean (left and middle panels), but differ considerably for the median (right panel). Furthermore, similar to previous results, the $t$ test and permutation test operating on the mean provide virtually the same power, and both are slightly higher than the power derived from the analytical and empirical distributions for intermediate $N$.

The lower panels of Fig. 3 depict type 1 (false rejection) error rates as a function of the sample size $N$ under $H_0$. The results

TABLE II

SUBJECTIVE (MUSHRA) SCORE SAMPLE SIZE, MEAN, STANDARD DEVIATION, SKEW AND KURTOSIS (PEAKEDNESS) FOR EACH CODEC SELECTED FROM THE MPEG USAC VERIFICATION TEST [26], [27]. THE COLUMN "ACRONYM" INDICATES THE ACRONYM USED THROUGHOUT THIS PAPER; THE COLUMN "MPEG TEST/SYSTEM" CONTAINS THE FULL TEST AND CONDITION NAME USED IN [27]. KURTOSIS IS EXPRESSED AS EXCESS KURTOSIS (E.G., THE KURTOSIS CALCULATED FROM THE DISTRIBUTION OF THE DATA MINUS THE KURTOSIS OF A GAUSSIAN DISTRIBUTION)

| Acronym | MPEG test/system | Size | Mean | Std | Skew | Excess kurtosis |
|---|---|---|---|---|---|---|
| USAC12M | Test 1 USAC 12M | 1656 | 62.6 | 17.1 | 0.2 | -0.4 |
| USAC16M | Test 1 USAC 16M | 1656 | 71.0 | 16.2 | -0.1 | -0.6 |
| USAC24M | Test 1 USAC 24M | 1656 | 79.0 | 14.4 | -0.5 | -0.1 |
| AMRWBP24S | Test 2 AMRWBP 24S | 1104 | 56.7 | 20.0 | 0.1 | -0.6 |
| USAC16S | Test 2 USAC 16S | 1104 | 63.3 | 17.1 | 0.1 | -0.5 |
| USAC20S | Test 2 USAC 20S | 1104 | 68.3 | 16.6 | -0.2 | -0.3 |

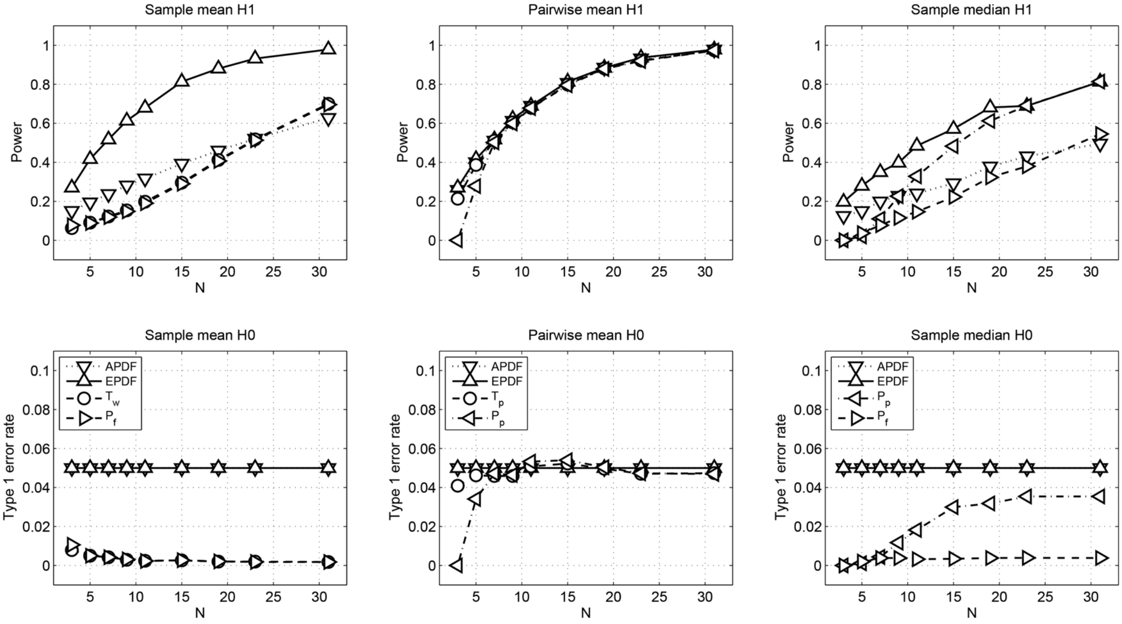

Fig. 4. Comparison of USAC12M vs USAC16M. Same layout as in Fig. 1.

for the student's $t$ test and permutation test using the mean are virtually equivalent and are both quite accurate in their rejection of $H_0$ given the nominal value of 0.05. The permutation tests employing the median, on the other hand, are typically more conservative while the type 1 (false rejection) error rates for two flavors of the permutation test depend on the sample size $N$ in a non-monotonic manner (leftward and rightward triangles in the bottom-right panel).

### B. MPEG USAC Verification Test

*1) Data Sets:* The MPEG Unified Speech and Audio Codec (USAC) was completed in 2011 and verification test results were made public in the same year [26], [27]. The verification test comprises three configurations that vary in bit rate and number of channels (mono or stereo). The tests that were used in the current analysis are listed in Table II. Test 1 is a mono test (12, 16 or 24 kbps), while test 2 focuses on stereo content (16, 20 or 24 kbps total). The test content consisted of 24 items containing speech, music or mixtures thereof. Table II lists the mean, standard deviation, skew and excess kurtosis for the MUSHRA data associated with each codec selected for this study. Except for differences between means, the higher moments are more consistent across codecs than those observed for the MPEG Surround verification test given in Table I.

*2) USAC12M vs USAC16M:* Test labels USAC12M and USAC16M represent scores for a mono configuration and have very similar population moments, except for a mean difference in subjective scores of 8.4 MUSHRA points. Results for power and type 1 (false rejection) error rates as a function of $N$ are shown in Fig. 4. When power based on a comparison of sample means is considered (top-left panel), the power derived from the empirical sampling distribution (upward triangles) is considerably higher than the power of the other methods. For pairwise comparisons using the mean as location statistic, on the other hand (middle-top panel), all methods perform virtually identically and generally provide the highest power compared to the sample mean or tests employing the median. For the later (right-top panel), the empirical distribution provides the highest power being slightly above the power associated with the pairwise permutation test (leftward triangles). The analytical PDF (downward triangles) and the permutation test using all permutations (rightward triangles) provide relatively low power.

Under $H_0$, as depicted by the lower panels in Fig. 4, comparison of sample means using a permutation or $t$ test results in quite conservative $H_0$ rejection rates, while for pairwise comparisons using the mean (bottom-middle panel), these rates are quite close to the nominal $\alpha$ of 0.05. The bottom-right panel denotes type 1 (false rejection) error rates for the sample median,

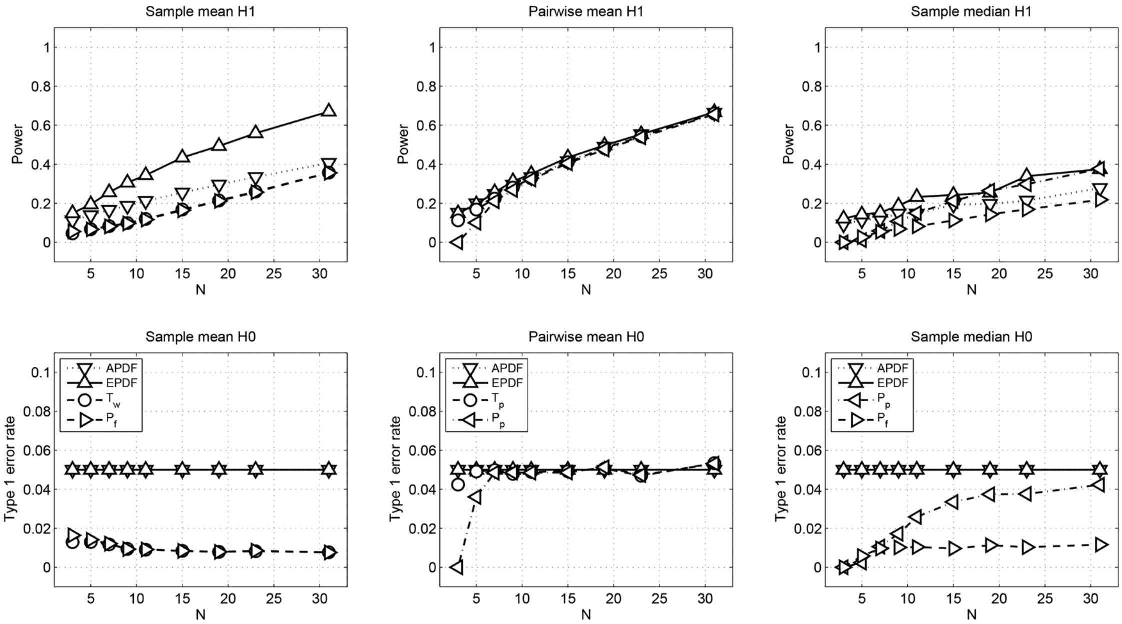

Fig. 5. Comparison of AMRWBP24S vs USAC16S. Same layout as in Fig. 1.

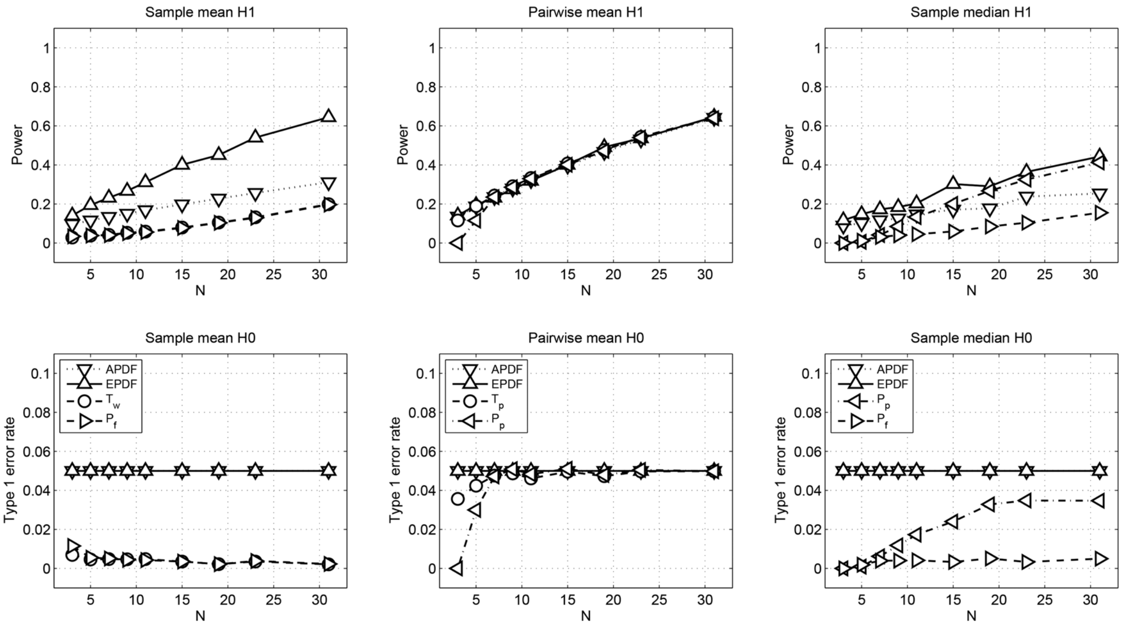

Fig. 6. Comparison of USAC16S vs USAC20S. Same layout as in Fig. 1.

showing conservative $H_0$ rejection rates for the full permutation test (rightward triangles), and somewhat more accurate rejection rates for the pairwise test (leftward triangles).

The exact same simulation was performed for a comparison between USAC16M and USAC24M. Despite minor qualitative differences, the results for that comparison are very similar to those obtained for USAC12M vs USAC16M, and are therefore not shown.

*3) AMRWBP24S vs USAC16S:* The results obtained for the comparison between AMRWBP24S vs USAC16S are shown in Fig. 5. When the power for a comparison of the sample means is considered (top-left panel), a similar trend is observed as in the previous simulations, in that power based on the empirical distribution of the mean (upward triangles) is highest, while the $t$ test (circles) and the permutation test (rightward triangles) are resulting in virtually identical power. For pairwise comparisons employing the mean, all methods are virtually identical (middle-top panel) and provide substantially higher power than for the permutation test with the median (right-top panel). The lower panels indicate the probability of a type 1 (false rejection) error as a function of $N$. All tests are conservative with respect to the nominal $\alpha = 0.05$ except for the methods employing the pairwise mean (bottom-middle panel).

*4) USAC16S vs USAC20S:* The results obtained for the comparison between USAC16S vs USAC20S are shown in Fig. 6. In this configuration, power as a function of $N$ for a comparison of sample means (top-left panel) is highest for the empirical distribution of the mean (upward triangles), followed by

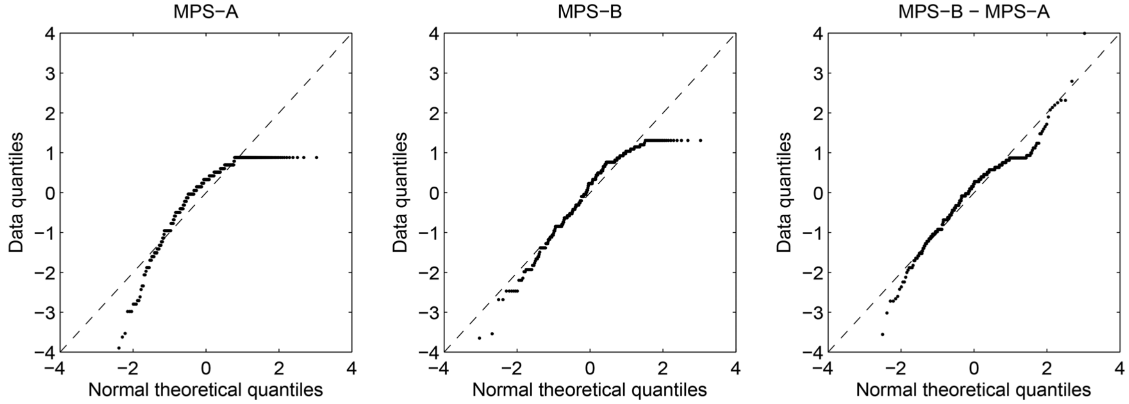

Fig. 7. Quantile-quantile plot of the distribution of the data against a theoretical normal distribution. The left, middle and right panels represent MPS-A, MPS-B, and pairwise differences thereof, respectively.

the analytical distribution of the mean (downward triangles). The permutation test and the $t$ test are virtually equivalent. For pairwise comparisons (top-middle panel), all methods perform very similarly and better than all tests using the sample median (right-top panel). The lower panels indicate the type 1 (false rejection) error rates showing results that are qualitatively in line with the results shown in the previous section.

## IV. Discussion

The Monte-Carlo simulations outlined in Section III cover a variety of spatial audio configurations (mono, stereo, and multichannel) as well as tests with different content types (music, speech, and combinations thereof). Despite these differences, several commonalities in the results across tested configurations can be observed. Firstly, the empirical distribution of the mean results in the highest power in almost all tested conditions, except for the comparison between MPS-A and REF. The empirical distribution of the mean is a special case because it provides the same result for a comparison of sample means and for pairwise differences, as in both cases, common linear factors such as overall differences among assessors or audio excerpts are canceled out due to the differencing operation, resulting in a relatively small variance in the difference between two means and hence resulting in high power. In fact, one could argue that the empirical distribution of the mean should be considered as the most accurate predictor for type 1 (false rejection) and type 2 (false nonrejection) errors in the traditional sense of statistical inferences based on an (assumed) distribution of the sample mean.

A second observation is that for a comparison of sample means, the analytical distribution of the mean performed worse (e.g., provided lower power) than the empirical distribution, while for pairwise comparisons, a very similar $H_0$ rejection rate between these methods is observed. The analytical distribution does not rely on normality nor does it impose requirements on any moment of the distribution, but it does assume independence in observations. The presence of any common factors in the data will increase the width of the analytical sampling distribution of the mean and therefore decrease power. If, on the other hand, these common factors are (virtually) identical for two systems under test, pairwise differences will provide transformed data that better satisfy the independence assumption. This effect is clearly observed in all tested conditions, as for pairwise differences, the power associated with the empirical and analytical distributions of the mean is virtually identical, while for tests on the sample mean, the empirical distribution provides higher power than the analytical distribution.

A third observation is the striking similarity between results obtained for the $t$ test, the analytical and the empirical distributions of the mean when pairwise comparisons are considered. This result indicates that for pairwise differences calculated from the data under test, the $t$ test provides very accurate results despite the fact that the data are not normally distributed. An example of such deviation from a normal distribution is depicted in Fig. 7. The figure shows a quantile-quantile plot of normal theoretical quantiles along the abscissa, with the ordinate indicating the quantiles of the actual data population after normalizing their mean and standard deviation. The different panels represent MPS-A, MPS-B, and their pairwise differences, for the left, middle and right panel, respectively. As can be observed from the figure, the deviation of the data quantiles away from the diagonal indicates that the assessor responses are not normally distributed, a finding in line with the 3rd and 4th moment (i.e., a negative skew and a positive kurtosis for MPS-A) listed in Table I. The pairwise difference distribution, depicted in the right panel, suggests a somewhat closer correspondence to a Gaussian distribution although the match is not perfect either. A likely explanation why these deviations from a normal distribution do not cause the $t$ test to produce inaccurate results is that the 95% critical value of the distribution of the pairwise *mean* is by virtue of the central limit theorem quite close to that of the $t$ distribution. The exception to this observation seems to be the comparison of MPS-A against the hidden reference (REF) depicted in Fig. 3. In that particular case, both the permutation test with the mean, as well as the $t$ test have a higher power than what is expected based on the actual sampling distribution of the mean. The likely cause of this observation is the significant skew present in the hidden reference data, resulting in a smaller critical value $x_t$ for the parametric $t$ test compared to the critical value determined from the actual sampling distribution of the mean. Such a too small critical value comes with the risk of an inflated type 1 (false rejection) error under $H_0$, but interestingly, this does not occur for the $t$ test nor for the permutation test based on the mean. This finding, in combination with the high kurtosis for the data at hand, suggest that the PDF of the mean is more peaky than a Gaussian distribution with the same standard deviation, preventing the type 1 (false rejection) error rate for a parametric test from becoming larger than the nominal 5% despite the optimistic critical value.

A fourth finding is that both the type 1 (false rejection) and type 2 (false non-rejection) errors associated with the parametric $t$ test and the non-parametric permutation test using the mean are very similar across all tested conditions. If the sample sizes of $N = 3$ and $N = 5$ are excluded, in which the $p$ value of the pairwise permutation test is lower bounded by the limited number of unique permutations (cf. Section II-D), the normalized correlation coefficient across all tested conditions amounts to 0.9997 (based on 2240000 $p$ value pairs). If binary decisions are correlated (e.g., rejections of $H_0$), a value of 0.9917 is obtained. Given the substantial differences in their approach, the underlying assumptions and the formulation of $H_0$ of these inference tests, this finding comes somewhat as a surprise. In particular, when applied to data that are not normally distributed, such as the aforementioned comparison of MPS-A against REF, the permutation test using the mean does not show any advantage or disadvantage in terms of power or type 1 (false rejection) error over the parametric $t$ test. The obvious benefits of the permutation test are that 1) the test does not rely on an assumed distribution of the data nor the distribution of a location statistic derived thereof, 2) it is typically simple to implement and 3) does not require calculation of critical values of a cumulative $t$ distribution. More specifically, when employing pairwise permutations exclusively and with the mean as location statistic, the accuracy of the permutation test was among the best performers of all methods tested, both in terms of type 1 (false rejection) error as well as providing high power.

A last but not least finding is the relatively conservative performance of the permutation test when employed with the *median* as sample location statistic. In virtually all conditions tested, the power of this particular test is lower than for corresponding methods using the mean. This finding suggests that at least for the data under test, the mean is a more robust parameter to predict whether $H_0$ is true or false. In particular, since this difference is also observed for the empirical distributions of the mean and median, the sampling distributions of the mean of $Y$ and $Z$ seem to be more distinct (e.g. having less overlap) than those using the median.

Although the applicability of non-parametric permutation test to subjective quality scores seems promising, there are a few aspects of this methodology that should be kept in mind. Firstly, it requires dedicated software modules to execute and analyze permuted data. Traditional statistical analysis packages may not all have the required functionality readily available. Secondly, if the data to be analyzed gives rise to a number of unique permutations that is too large to evaluate in an exact test, sampling permutations inherently introduces some randomness in the resulting $p$ value and its outcome depends on the permutation sampling algorithm. Although the experimenter can control this variability by adjusting the number of permutations, it does add a layer of conceptual complexity having to consider confidence intervals of probabilities. Lastly, the computational complexity of a permutation test is typically significantly larger than that of a parametric test, but with today's abundant availability of computing power this disadvantage is typically of little practical relevance.

### A. Limitations

The results obtained in this study cannot be generalized to arbitrary subjective quality data. Differences in test excerpts, audio codecs, assessors, listening environments or testing protocols may give rise to dissimilar statistical structures in the data and hence the various statistical inference tests may perform differently. Moreover, only pairwise comparisons were evaluated, ignoring the potential presence and effect of covariates or higher-order interactions. The Null hypothesis was simulated by equalizing means and subsequent pooling to result in identical PDFs of the two codecs under test. Any modifications in this procedure (for example by only equalizing a certain measure of central tendency while preserving differences in higher-order moments) may give rise to different results and conclusions. The use of the arithmetic mean as a metric for central tendency inherently assumes an interval or ratio scale for subjective quality ratings which may not always be justifiable (cf. [28]).

## V. Conclusions

Monte-Carlo simulations of statistical inference tests were performed to assess type 1 (false rejection) and type 2 (false non-rejection) errors associated with subjective audio quality data for small sample sizes (up to 31). The results indicate that pairwise comparisons are essential if high power is desired. Such pairwise comparisons provide (partial) cancellation of variance due to common factors such as excerpt and assessor dependencies. These pairwise inferences can be realized by a parametric $t$ test or by a non-parametric permutation test with the mean as location statistic, provided that for the latter, only pairwise permutations are evaluated. For such pairwise comparisons, the $t$ test and the permutation test showed a remarkable correspondence in their $p$ values and are in line with data obtained using empirical data and analytical expressions of the distribution of the mean, indicating that despite the data deviating from a normal distribution, both tests resulted in predictions of type 1 (false rejection) and type 2 (false non-rejection) error probabilities without bias. The median as a sample location statistic in combination with the permutation test performed comparatively conservatively, both in terms of the type 1 (false rejection) error rate and power. Although one should be extremely careful not to generalize the conclusions from the results obtained for the specific data sets used in this study, the data do suggest that a resampling test using pairwise permutations and the mean as location statistic is an interesting candidate for further study. Specifically, the absence of any requirements on data distributions, the close match between the actual and nominal type 1 (false rejection) error rate, and its relatively high power seem valuable properties for the statistical analysis of subjective audio quality data.

## Acknowledgment

The author would like to thank the anonymous reviewers and the associate editor for their insightful and helpful comments to improve the manuscript.

## REFERENCES

[1] *Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)*, ITU-R Rec. BS.1534-1, Int. Telecomm. Union, 2003.

[2] *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, ITU-R Rec. BS.1534-2, Int. Telecomm. Union, 2014.

[3] L. Leventhal, "Type 1 and type 2 errors in the statistical analysis of listening tests," *J. Audio Eng. Soc*, vol. 34, no. 6, pp. 437–453, 1986 [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=5265

[4] F. Nagel, T. Sporer, and P. Sedlmeier, "Toward a statistically well-grounded evaluation of listening tests - avoiding pitfalls, misuse, and misconceptions," in *Convention paper 8146 128th AES Conv.*. London, U.K.: Audio Engineering Society, May 22–25, 2010.

[5] C. A. Boneau, "The effects of violations of assumptions underlying the t test," *Psychol. Bull.*, vol. 57, no. 1, p. 49, 1960.

[6] G. E. P. Box and G. S. Watson, "Robustness to non-normality of regression tests," *Biometrika*, no. 1/2, pp. 63–106, Jun. 1962.

[7] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey*, Robust Estimates of Location: Survey and Advances*. Princeton, NJ, USA: Princeton Univ. Press, 1972.

[8] G. E. P. Box, G. C. Tiao, C. W. J. Granger, and I. Guttman*, The collected works of George EP Box*. Monterey, CA, USA: Wadsworth, 1985.

[9] E. J. G. Pitman, "Significance tests which may be applied to samples from any populations," *J. R. Statist. Soc.*, vol. 4, no. 1, pp. 119–130, 1937.

[10] M. D. Ernst, "Permutation methods: A basis for exact inference," *Statist. Sci.*, vol. 19, no. 4, pp. 676–685, 2004.

[11] R. B. Anderson, "Conceptual distinction between the critical p value and the type I error rate in permutation testing," *J. Modern Appl. Statist. Meth.*, vol. 12, no. 1, pp. 2–8, May 2013.

[12] J. P. Romano, "On the behavior of randomization tests without a group invariance assumption," *J. Amer. Statist. Assoc.*, vol. 85, no. 411, pp. 686–692, 1990.

[13] A. F. Hayes, "Randomization tests and the equality of variance assumption when comparing group means," *Animal Behaviour*, pp. 653–656, 2000.

[14] Y. Huang, H. Xu, V. Calian, and J. C. Hsu, "To permute or not to permute," *Bioinformatics*, vol. 22, no. 18, pp. 2244–2248, 2006.

[15] M. Aickin, "Invalid permutation tests," *Int. J. Math. Math. Sci.*, pp. 1–10, 2010, article ID 769780.

[16] A. M. Mood, F. A. Graybill, and D. C. Boes*, Introduction to the theory of statistics*, 3rd ed. New York, NY, USA: McGraw-Hill, Jun. 1974.

[17] J. F. Kenney and E. S. Keeping*, Mathematics of Statistics*, 3rd ed. Princeton, NJ, USA: Van Nostrand, 1962, vol. Part 1, ch. The Median, pp. 211–212.

[18] J. Cohen, "The earth is round ($p < .05$)," *Amer. Psychol.*, vol. 49, pp. 997–1003, 1994.

[19] G. Gigerenzer, S. Krauss, and O. Vitoch, "The Null ritual: What you always wanted to know about significance testing but were afraid to ask," in *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA, USA: Sage, 2004, pp. 391–408.

[20] F. E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics Bulletin*, vol. 2, no. 6, pp. 110–114, Dec. 1946.

[21] P. R. Peres-Neto and J. D. Olden, "Assessing the robustness of randomization tests: Examples from behavioural studies," *Animal Behaviour*, vol. 61, pp. 79–86, 2001.

[22] C. Ninness, R. Newton, J. Saxon, R. Rumph, A. Bradfield, C. Harrison, and E. Vasquez, III, "Small group statistics: A Monte Carlo comparison of parametric and randomization tests," *Behavior Soc. Iss.*, vol. 12, pp. 53–63, 2002.

[23] "N8851 report on MPEG surround verification tests," ISO/IEC JTC1/SC29/WG11 (MPEG), Jan.–May 2007–2014 [Online]. Available: http://mpeg.chiariglione.org/standards/mpeg-d/mpeg-surround/report-mpeg-surround-verification-tests

[24] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par, "Background, concept, and architecture for the recent MPEG surround standard on multichannel audio compression," *J. Audio Eng. Soc.*, vol. 55, no. 5, pp. 331–351, 2007.

[25] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and D. S. Chong, "MPEG surround - the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.

[26] S. Quackenbush and R. Lefebvre, "Performance of MPEG Unified Speech and Audio Coding," in *Convention paper 8514 131th AES Conv.*. New York, NY, USA: Audio Eng. Soc., Oct. 20–23, 2011.

[27] "N12232 unified speech and audio coding verification test report," ISO/IEC JTC1/SC29/WG11 (MPEG), May–May 2013–2014 [Online]. Available: http://mpeg.chiariglione.org/standards/mpeg-d/unified-speech-and-audio-coding/unified-speech-and-audio-coding-verification-test

[28] S. Zieliski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests - a review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, Jun. 2008.

**Jeroen Breebaart** received an M.Sc. degree in biomedical engineering from the Eindhoven University of Technology in 1997, and a Ph.D. degree in psychophysics from the same university in 2001. From 2001 to 2007, he was with the Digital Signal Processing group at Philips Research, conducting research in the areas of spatial perception, stereo and multi-channel parametric audio coding, automatic audio content analysis and binaural rendering, and was involved in the development of several international standards, such as MPEG enhanced aacPlus, MPEG Surround, and Spatial Audio Object Coding (SAOC). From 2007 to 2010, he was with the Information and System security group at Philips Research, and worked in the areas of biometric information protection and automated human behavior analysis. He actively participated in the ISO/IEC IT security techniques standardization committee (JTC1 SC27) as co-editor of project 24745 (Biometric information protection). He also participated in the ISO/IEC Biometrics standardization committee (JTC1 SC37), and was involved in several EU-funded projects (3D FACE and TURBINE). In 2011 and 2012, he was with Civolution, developing watermarking algorithms for broadcast monitoring, second screen applications and forensic tracking of media content. In 2012, he moved to Sydney, Australia, to join Dolby Laboratories.