USING STATISTICS TO ANALYSE LISTENING TEST DATA: SOME SOURCES AND ADVICE FOR NON-STATISTICIANS

L E Harris ISVR, University of Southampton
Email: lh1@isvr.soton.ac.uk
K R Holland ISVR, University of Southampton

1 INTRODUCTION

Subjective listening tests are still an extremely useful way to assess the quality of audio reproduction equipment. It is now accepted that such tests can be regarded as scientific experiments if the test procedure and conditions are very carefully planned and controlled. Much has been published on this topic, and guidelines for carrying out rigorous tests can be found quite easily by those wanting to conduct their own. However, advice on how to analyse data from the experiments is less accessible within audio literature, despite this being a crucial part of the research process. Any careful preparation for the listening tests will be wasted if the results are poorly analysed and incorrect conclusions about the data are drawn.

Section 2 of this paper contains a short review of some sources that may be useful to those who have little or no statistical knowledge but want to carry out a basic analysis of their own data. *Section 3* describes a case study where several different 'virtual' loudspeakers were compared in an ABX listening test; the statistical analysis applied to the data is briefly explained and demonstrated in *section 4*. Due to the type of data yielded by such tests, the paper focuses on non-parametric analysis; this is encountered much less often in audio literature than parametric methods (e.g. ANOVA) despite ABX comparisons being a popular testing strategy when evaluating audio devices.

2 SOME USEFUL SOURCES OF STATISTICAL INFORMATION

Presented here are some journal papers and books that contain a lot of helpful information for those planning listening tests with statistical analysis in mind. *Section 2.1* focuses on sources easily found within the audio literature; *section 2.2* has details of useful non-audio texts, many of which were written for students of the social sciences.

2.1 Audio Literature

A good starting point is the 'Great Debate' paper by Lipshitz[1]. Though this does not contain any theory or specific examples of statistical analysis, it addresses several fundamental concepts, such as randomisation and significance. Geddes[2] also raises some interesting general points for anyone planning listening tests of the ‘paired comparison with reference’ type, though the title refers specifically to compression drivers.

When planning listening tests that generate rank data, EBU Tech. 3286[3] may be helpful in the early stages as it briefly mentions non-parametric analysis and why this should be used. However, it gives very little further statistical explanation or advice on processing data of this type. Its

description of the 'distribution profile' does not show it to be very useful in a statistical sense, and is also slightly misleading: in *Appendix L*, the profiles are referred to as bar charts, but from inspection alone of the diagram in *Appendix I*, it would appear that the ranking categories are a continuous linear scale. To be strictly correct, the bars should be separated to show that they are discrete categories[4].

ITUR BS116-1[5] is a more substantial document, and *Annex 1*, particularly *sections 1 – 6*, give clear and concise guidelines on planning effective listening tests. *Sections 9 and 10* then give useful information on analysis of the subsequent tests, with *appendices 1 and 2* having a detailed and interesting account of methods to assess listener expertise. However, This Recommendation is solely concerned with parametric analysis as the results are given on a continuous grading scale, though it does make a brief reference to non-parametric data in *section 9*; here the important point is highlighted that unless the test grading scale can be shown to be linear, comparisons of different grades can only be considered as ranks.

Bech's paper on evaluating data from listening tests[6] provides a good generalised overview of statistical analysis, and *section 5* briefly discusses parametric and non parametric techniques. However, it is probably most useful for those working with parametric methods such as ANOVA (i.e. using data that has been generated on a linear grading scale). *Perceptual Audio Evaluation*[7] by the same author is an excellent and comprehensive book for information on all aspects of subjective testing; unfortunately, much of the chapter on statistics is quite complicated to understand for statistical newcomers. It also lacks a detailed description of non-parametric methods (though it does not dismiss their use and cites some recommended references). For this reason, it is suggested that readers look to other sources first for guidance.

Leventhal[8] is an excellent source for learning about key statistical concepts and how to apply them in an audio context i.e. listening tests. It focuses on binomial experiments (see *section 4* for more information) and hence, non-parametric data analysis methods. It also provides reference tables of pre-calculated error probabilities in this type of experiment for a range of sample sizes. In *section 4*, the interesting concept of a Fairness Coefficient is introduced and demonstrated; this may be best understood once the reader already has a good grasp of type 1 and type 2 error testing. Overall, the paper contains a lot of important statistical information that may take several attempts to fully comprehend. It is strongly advised that readers return to this paper before and after familiarising themselves (from other sources) with the topics it covers. Following on from this directly is a paper by Burstein[9]; this is extremely useful for those with sample sizes other than those listed by Leventhal. Several formulae are presented that are based on an approximation to the normal distribution. This vastly simplifies significance testing on binomial data for all but very small samples sizes (below 15) by replacing very cumbersome calculations with a simple equation. In addition to these papers, Leventhal and Huynh[10] present an interesting description and examples of directional significance testing in audio experiments. Some of its content will be helpful to those of any statistical ability, but a good understanding of the basic concepts involved is required to fully make use of the information in this paper.

2.2 Other Disciplines

For a clear, concise, and well-presented text, readers are strongly advised to see the book by Argyrous[4]. This combines a thorough description of statistical methods from the most basic level upwards; though it will be especially helpful for those using the computer program SPSS, the well-explained examples and logical progression through topics will be useful to any student of this subject. *Modern Elementary Statistics* by Freund[11] is also well presented with key points being well highlighted in every chapter. It covers a wide range of statistical topics and will appeal to those who like to learn through exercises and examples.

The book by Meddis[12] focuses on rank analysis of data. *Chapters 1-5* are informative and easy to read, and make very good background reading for anyone new to the subject of nonparametric statistics. From *chapter 6* onwards the focus is on examples and computation; it becomes increasingly complicated and it is difficult to match the examples with typical listening test scenarios, so this text may be of limited use beyond the first five chapters. Another book focussing only on nonparametric statistics is that by Siegel[13]; the topic is given extensive coverage and as such, some complex methods are described. *Chapters 1-3* provide a useful introduction to statistical analysis, though does contain some mathematics. Despite being dense with information, and therefore sometimes difficult to follow, this is a good reference book with detailed descriptions and examples of each technique.

For those wanting a more literary introduction to statistical concepts, *Statistics Without Tears* by Rowntree[14] is a highly readable and relatively short text. It gives an overview of the fundamental topics that must be understood before any analysis of 'real' data can be done, and contains very little maths. This is recommended as additional background reading rather than as reference for specific examples.

2.3 Software

There are several widely available computer programs that can perform statistical analysis, though it is not advisable to use any of them without a basic understanding of the functions they perform and the large number of options they offer. Perhaps the most common program is Microsoft Excel. This makes data entry and manipulation very easy and will perform many statistical functions. MATLAB has a statistical toolbox and good help files to accompany the functions; it is also quite difficult to use without a very good understanding of the statistical processing it performs, even for those already familiar with the program.

This author found SPSS a very useful statistical package, following a basic introductory tutorial. Visually, it is much like Excel and contains a vast array of processing options for all kinds of data. It can be used effectively with a fairly basic understanding of statistics and also has good help files and tutorials for new users. The book by Field[15], aimed at degree undergraduate students, is light-hearted and easy to read; it has a huge range of specific problems, examples, and step-by-step guides to solving them using SPSS, accompanied by concise background theory on each technique.

3 CASE STUDY OF AN 'ABX' LISTENING TEST: EXPERIMENT DETAILS

When planning an experiment, it is important to consider what kind of analysis will be applied to the data. This section describes a set of listening tests that were designed considering several important features: minimising all sources of bias and controlling 'nuisance variables', maintaining independence between trials (so the result of any trial does not influence the result of the others), and generating data of a type that was suitable for statistical analysis after the experiments. These factors would determine whether useful analysis could be performed and ultimately, whether it was worthwhile performing the listening tests.

3.1 Design And Execution

Sections 3.1.1 to *3.1.3* describe why the experiment was carried out, how it was carried out, and some details on the software written to control the tests. Justification for each design feature is not included in the text, but most features were implemented after an extensive literature search on subjective audio testing and statistical analysis methods for experimental data.

3.1.1 Motivation And Aims

It is generally accepted within the audio industry that a flat and extended frequency response signifies a high performance loudspeaker. An accurate response is not just desirable, but essential in the case of studio monitors; here the engineer must be presented with a realistic impression of what is, or has been, recorded in order to make appropriate changes to relative levels of instruments within the overall mix. It has been demonstrated that some monitors, though appearing to be very high fidelity when viewed solely in the frequency domain, exhibit poor time response performance, having a characteristic decaying 'tail' in the bass region. This ringing leads to loss of musicality, sometimes known as 'one note bass', where timing becomes blurred, and key instruments with fundamental frequencies in that range (primarily kick drum and bass guitar) are at risk of being incorrectly balanced.

The experiments were part of a project to investigate a novel method of measuring low frequency quality in loudspeakers, primarily targeted at reproduction of music in professional applications (i.e. studio monitoring). The technique aims to account for transient behaviour, or time response performance, and is based around the Modulation Transfer Function (MTF); this is normally used in acoustics as the basis of the Speech Transmission Index for gauging speech intelligibility inside listening spaces. The aim of the listening tests was to try and establish whether the new technique is a good indicator of subjective bass accuracy.

3.1.2 Experimental Procedure

The experiment was designed to make subtle differences between loudspeakers as audible as possible: Monophonic listening tests (a single loudspeaker), with individual subjects were performed in the large ISVR[*] anechoic chamber with the door shut. Deconvolution was used to equalise the low frequency response of a large recording studio monitor. This gave a frequency response that was perfectly flat in the bass region (0dB below approx. 100Hz), and totally unaltered in the mid and high frequencies. A MATLAB function was written to generate 'woofer models'- using five of the Thiele-Small parameters commonly found on manufacturers' data sheets, the complex frequency response could be simulated for a given loudspeaker driver. The function allowed various design features to be selected, most notably cabinet volume, whether the cabinet was sealed or ported, and whether or not a low-frequency protection filter was to be used.

Using parameters of real drivers, a number of 'woofer models' were created and compared. Five were selected that gave a good range of bass response, differing in their low frequency -3dB point and order of roll-off i.e. the characteristic response variations between loudspeakers with sealed and ported cabinets. A 'Reference' model response was also generated using the transfer function equation for a high pass filter. Thus, six 'target bass responses' were created for comparison in the listening tests. The responses were then imposed onto the equalised-flat monitor response; when music was replayed through these models, the listener would effectively be hearing six loudspeakers identical in every way except for their low frequency behaviour.

[*] Institute of Sound and Vibration Research, University of Southampton

The test source signals were 25s extracts from commercially available musical recordings. To reduce the possible effects of programme dependence (see *section 4.3*), three extracts were chosen. These were carefully selected according to a list of criteria, including adequate energy at low frequencies i.e. bass content. Before any processing, the RMS values of the extracts were matched. Then the extracts were listened to under test conditions and adjusted as necessary so that they had approximately equal loudness. Level differences between designs *within each extract* due only to the difference in bass response were preserved.

The experimental loudspeaker was raised on a stand so that the mid/high driver was approximately at ear level of the listener. The listener was seated on-axis in a comfortable chair at a distance of 2m from the front of the loudspeaker cabinet. The test was fully automated by computer and controlled by the listener through a small touch-screen interface. Sound pressure level, measured at the listening position with a calibrated sound level meter, was approximately 76dB LAeq for each extract and was not altered after set-up.

The test was a 2-Alternative Forced Choice design[16]; listeners were required to say which of two loudspeaker models, randomly assigned to A and B, sounded *most like* the Reference, X. It was made clear prior to the tests that they should not answer based on personal preference, but they were not told anything about the nature of the differences between the designs. Five loudspeaker designs, plus a 'hidden Reference' were tested. At the start of each trial, one of the musical extracts would play; the listener could switch freely between the three channels (i.e. three loudspeaker designs) using a 3-way switchbox. They would then record their answer by pressing 'A' or 'B' on screen and move onto the next trial.

The test was double blind and fully randomised using a MATLAB function, written to create an individual 'playlist' of files for each participant. The software automatically split the playlists into two sessions so that maximum session duration was less than 30 minutes to prevent listener fatigue. For each trial, the function randomly assigned: which pair of designs would be compared, which extract would be used, and which channel (A or B) each of the designs would play back through. It also ensured that the same extract would never be used for two consecutive trials.

Adapting the equation given in McCormick[17], the number of trials, N_t, can be calculated as:

$$N_t = \left(\frac{M(M-1)}{2} \right) N \qquad (1)$$

Where: M = *number of designs* (6) and N = *number of extracts* (3), giving (15 x 3) = 45 trials for this test.

Anyone wanting to take part was allowed to do so unless they knew of any physical problems seriously affecting their hearing. Participants were asked to complete a short questionnaire on their listening experience and general health (such as any recent colds or ear infections). Listeners were then provided with written and verbal instructions before the test but there was no formal training as they were already being asked to attend two sessions without payment. As a compromise, three warm-up trials were carried out immediately before the start of each listener's first session. This let them practice controlling the test and hear the three extracts that would be used in the actual experiment. In addition, three 'blind trials' or dummy-runs were included at the start of each session to let the listener acclimatise a little to the task; two more were also added at the end of each session, making 55 trials in total. The dummy trials were identical for all listeners (non-randomised) and were included in the formal testing sessions; listeners were not aware that they were dummy trials.

3.1.3 Testing Software

A graphical user interface (GUI) was written in MATLAB that listeners used to control replay during the test. Through randomisation and this removal of the need for experimenter-interaction, it was ensured that the tests were double blind and generated independent data. Visually, the interface was very simple, but it performed a number of important and sometimes complicated functions.

A screen shot of the main test screen is shown in *figure 1*. The buttons were made to be large, clear and well separated on the 8 inch touch-screen tablet, minimising the risk of accidentally pressing the wrong button. Neutral colours were used for the 'A' and 'B' buttons so as not to subconsciously bias the listener into choosing one or other, or thinking that one was 'right' and the other 'wrong'. The interface was programmed so that if the A or B buttons were pressed before 'PLAY', no answer would be registered and the test would not move on; this prevented answering without listening to the extract in each trial.

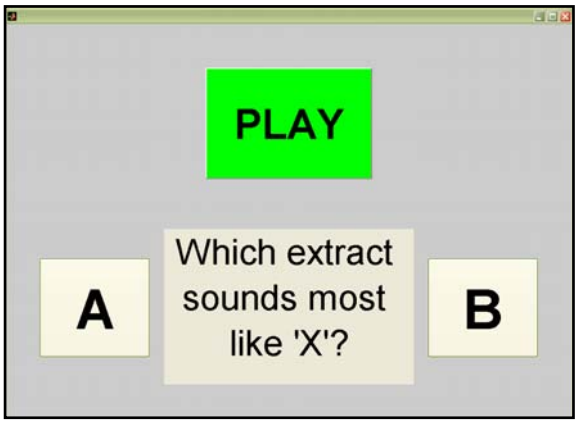

Fig 1. MATLAB GUI example screenshot

As mentioned in *section 3.1.2*, a separate MATLAB function used prior to the test would generate a new, fully randomised playlist of stimuli for each listener entered onto the system; this was labelled with their unique 'ID Tag' (initials and a number, automatically incremented if those initials had already been entered). This was automatically written to an Excel spreadsheet. During testing, the listener's ID Tag was selected from a drop-down list on screen; the software would then look up the appropriate playlist from the spreadsheet and play the relevant stimulus. All files were stored on the hard disk of a desktop computer and called up as necessary by the test program.

When the listener registered their answer in a trial, the result would be recorded into a separate 'results' sheet, along with all other relevant details. In this way, all results were automatically stored in electronic form, immediately ready for manipulation or processing. The onerous data entry stage required with paper results sheets was avoided, along with the possible risk of making mistakes when manually entering vast amounts of data.

Additional note for anyone using the 'rand' function in MATLAB: unless the state is set when calling this function, the same set of random numbers will be produced in every new session of MATLAB. Hence, if this had not been set in the random playlist program described above, new listeners' playlists would have been duplicated every time their IDs were registered after restarting the computer (e.g. At the start of each day of tests).

3.2 Handling The Data

23 listeners took part in the experiment. Along with the answer given by listeners, other data was automatically written into the results spreadsheet in a separate entry for each trial: date, session number, listener ID, which pair of loudspeaker designs was evaluated, and in which order they were assigned to channels A and B. This information was pasted directly into the statistical package SPSS. The data was then manipulated as required to perform each stage of the analysis. The five 'dummy runs' from each listening session were excluded from analysis using the filtering feature in SPSS. More specific details of the processing are given in *section 4*.

4 CASE STUDY OF AN 'ABX' LISTENING TEST: STATISTICAL ANALYSIS

Statistics cannot be used to prove or disprove a hypothesis, but they can be used to support or dispute one. If an appropriate test is used correctly, statistical methods can be very reliable, both in describing features of a data set and predicting things from it. It is up to the experimenter to choose an appropriate test; this will depend on many things, including what form the data is in. The results from an ABX test are basically frequencies: counts of the number of times A or B is selected.

For the experiment described in *section 3*, the primary data set was the number of times each loudspeaker design was chosen over another; from this, a 'subjective accuracy' rank order for the designs could be made. For many listening tests, the answers are given on a continuous, linear rating scale and parametric analysis can be used. If this is the case, the data is assumed to be from a normal distribution and hypotheses about a population with this distribution can be tested; calculation of a mean score is perhaps the most basic test of this type. For rank data, or data arranged into categories, it cannot be assumed that the data is sampled from a normally distributed population. As such, non-parametric techniques must be used; these do not make assumptions about the population distribution and are often considered to be 'short-cut' statistical methods, inferior to parametric equivalents. Bech[6] demonstrates that this is not necessarily the case, and explains why ordinal rating scales (rankings) can be preferable in listening tests to interval ones (numerical ratings). This is the reason why a quality ranking was sought in the case study: it was anticipated that most of the listeners would be inexperienced and thus, unable to consistently and accurately give quality scores on an interval scale. Though this meant that the data could not show *how much* better each design was than another, it was felt that demonstrating an effect was the priority; the removal of potential bias from incorrect use of rating scales by non-expert listeners was preferable to gathering extra data that might be useless.

It is important to note an inherent assumption in the experimental method: listeners were asked to give an answer based on which of the two designs sounded closest to the Reference design. This Reference was a model considered to be of a higher quality (more accurate at reproducing music) than any of the others in the test. Thus, if a listener judged a particular design as sounding closer to the Reference, it was implied that it was of a higher quality than the other design in that trial.

4.1 Calculating Rankings From Frequency Counts Using The Binomial Distribution

This section describes how the loudspeaker designs were ranked after the listening tests. Initially, the raw data was a long list of A's and B's; this was processed so that it was clear what each 'A' and 'B' meant: which pair of designs was being compared, with which extract, and which of them the listener chose as sounding most like the Reference. *Sections 4.1.1* and *4.1.2* explain how these processed results were then analysed.

4.1.1 Count Data And The Binomial Distribution

Data that can only fall into one of two categories is called binomial, or dichotomous. Classic examples are: Heads *or* Tails, Yes *or* No, Male *or* Female. The results from an experiment like this will be the counts, or frequencies, that occur in each category. In any sample, some random variation (sampling error) can be expected. For example, 10 coin tosses will not always return 5 heads and 5 tails, though in an infinite number of trials (the 'population') they would be 50% heads and 50% tails because the probability of getting either one in any trial is identical (0.5). So, for their sample of data, the experimenter needs to know how likely it is that variations between frequencies

in each category are due to chance (random variation) rather than some experimental effect. Inference tests i.e. predictions about the population from which that sample of results came, can be conducted to find out the probability of variations in the data being due only to chance, and therefore, how likely it is that those variations are actually because of the experimental effect under investigation. For small samples, the exact binomial distribution should be used to calculate probabilities for this.

As shown by Leventhal[8] (who explains binomial experiments clearly in this paper), calculations for the exact binomial distribution become very lengthy beyond tiny sample sizes. Fortunately, for larger samples[†], a simple formula can be used instead. This is based on an approximation of the binomial distribution to a normal one, thereby allowing a standard *z-test* to be carried out (details of the *z-test* will be found in the early chapters of any statistical textbook). In the case study, this method was used to analyse the split of listeners' answers between each pair of loudspeaker designs: were the relative percentages different enough to be able to conclude that variations between the designs were genuinely and consistently audible? Or were the listeners, not able to choose one design over the other, voting randomly between A and B, therefore bringing the results out to be roughly (but not exactly), 50-50? The method and results are given in *section 4.1.2*.

4.1.2 Ranking The Loudspeaker Designs

The A/B results were collated for each pair of loudspeakers compared in the test, regardless of which extract and channel assignment had been used. For each pair, the sample size was 69; the normal approximation to the binomial distribution could therefore be used. The counts were converted into relative percentages i.e. the percentage of the results *for that pair* that each design was awarded. From Argyrous[4], *the z-score* for a binomial percentage (when using the normal approximation) is:

$$z_{sample} = \frac{(P_s - 0.5) - P_u}{\sqrt{\frac{P_u(100 - P_u)}{n}}} \qquad (2)$$

When $P_s > P_u$, where P_s = *sample percentage*, P_u = *assumed population percentage*, and n = *sample size.*

In this case, the null hypothesis, H_0, is that the listeners are voting randomly because they cannot choose one design over the other. Thus, H_0: P_s = 50%. The alternative hypothesis, H_A, is that one of the designs is consistently chosen as being audibly closer to the reference. Thus H_A: P_s > 50%. Note that this is a *directional* alternative hypothesis; therefore, the test will be one-tailed, and $P_s > P_u$. See Leventhal[8,10] for an explanation of directional testing.

The individual *p-values* (probabilities) for each pair of percentages can be calculated by first finding the *z-scores* and then looking them up in a table of critical values for the normal distribution. Alternatively, the question could be asked: how much higher than 50% does the score percentage have to be to safely assume that listeners were voting due to genuine audible differences rather than chance? Rearranging equation (2) gives:

$$P_s = z_{sample}\sqrt{\frac{P_u(100 - P_u)}{n}} + P_u + 0.5 \qquad (3)$$

[†] The values vary slightly between sources, but Argyrous[4] quotes N>30 when working with percentages. Burstein[9] stated N>15 in his approximations using integers.

An appropriate value for *z* must be chosen; there is no correct value for this, and the standard significance level of 5% was used: *p = 0.05* one-tailed*, z = 1.64*. Therefore:

$$P_s = 1.64\sqrt{\frac{50(100-50)}{69}} + 50 + 0.5$$

Thus, P_s = 60.4 (1 d.p.)

So in any pair, the design with 60.4% or more of the result can be considered as the one audibly most like the Reference, and therefore the one more accurate at reproducing music. At this significance level, the probability of getting this distribution of answers by chance is, at the most, 5 in 100. *Table 1* lists each pair of loudspeaker designs and the frequencies given by listeners. Columns *d1* and *d2* are *design 1* and *design 2* in each pair respectively; e.g. for the first pair, *d1* is design R and *d2* is design C. The next two columns are the corresponding relative percentages. *Grey shaded pairs are not significant at the 5% level, one-tailed.*

A/B Pair	d1	d2	d1%	d2%
RC	46	23	66.7	33.3
RD	66	3	95.7	4.3
RE	66	3	95.7	4.3
RF	39	30	56.5	43.5
RG	57	12	82.6	17.4
CD	68	1	98.6	1.4
CE	61	8	88.4	11.6
CF	32	37	46.4	53.6
CG	36	33	52.2	47.8
DE	2	67	2.9	97.1
DF	2	67	2.9	97.1
DG	4	65	5.8	94.2
EF	10	59	14.5	85.5
EG	23	46	33.3	66.7
FG	49	20	71.0	29.0

Table 1. Distribution of subjective votes for each pair of loudspeakers compared in the listening test

The ambiguity of the three pairs in which neither design was clearly superior, meant that five rankings were possible at this significance level:

1) R F C G E D
2) R F G C E D
3) R C F G E D
4) F R C G E D
5) F R G C E D

Note that *4)* and *5)* are conspicuous in that 'F' is ranked closer to the Reference than the hidden Reference; this suggests that F was not audibly distinguishable from the reference to most listeners.

4.2 Assessing Listeners' Performance Using The Binomial Distribution

Ideally, each listener would have repeated the experiment several times so that their individual ability to give consistent answers could be assessed ('intra-listener reliability'). This was not practically possible, but the hidden Reference was built into the test as a compromise: without the listeners knowing, one of the designs to be evaluated in the test was the Reference itself. The idea behind using this method is that if a listener did not choose e.g. A, when A is identical to X but B isn't, they probably aren't able to perform the rest of the rest of the experiment properly. As described in *section 4.2.1*, the results from the hidden Reference comparisons were used to hypothesise about each listener's performance in the rest of the experiment.

4.2.1 Using The Hidden Reference

Despite the listeners being clearly informed that personal preference should not influence their answers, it must be inevitable that subjective bias will have been imposed on the judgements in

some trials when choosing between A and B. However, in trials containing the hidden reference, there could be no subjectivity: a listener's answer was either right or wrong.

A key assumption in using this hidden reference method is that the listener will complete those particular trials with roughly the same amount of (or lack of) subjectivity that they do for all the others. Though this cannot be assured, the use of double blind testing and full randomisation of stimuli playback will have reduced the effects of his assumption not being true.

As explained in *sections 4.2.2* and *4.2.3*, the results of this listener-performance analysis was not only interesting, but also useful for reducing some of the ambiguities in the overall loudspeaker rankings. It was concluded that this justified the 'high cost' of the method in terms of how many trials it required from the overall experiment (that could otherwise have been used in comparing more experimental designs).

4.2.2 Ranking The Listeners

The listeners were ranked by the number of times they correctly identified the hidden Reference; this appeared in five pairs and was rated by each listener with each extract. Therefore, sample size, n = (3x5) = 15. This value was not comfortably high enough to use the normal approximation for the binomial percentage that featured in *section 4.1*., so the exact binomial distribution was used. A two-tailed test was chosen this time because, as explained by Leventhal and Huynh[10], if listeners perform *worse* than chance this can indicate a defect in the experimental method. The hidden Reference was an appropriate case for investigating this possibility because, unlike the other trials, the listeners would definitely be either right or wrong in the absence of experimental faults.

This time there are two alternative hypotheses as well as the null that listeners answer by chance (H_0: p = 0.5):

H_{A1}: Listeners perform better than chance (identify the hidden ref. in most cases), $p > 0.5$
H_{A2}: Listeners perform worse than chance (possible experimental defect), $p < 0.5$

Referring to the table for a directional two-tailed test[10], at the 5% significance level, listeners must identify the hidden reference at least 13 out of 15 times to reject H_0 in favour of H_{A1}. If they fail to identify it at least 4 times, H_0 can be rejected in favour of H_{A2}.

Looking at the listener ranking, 10 out of 23 participants identified the hidden Reference at least 13 times. The lowest performing listener identified it 7 times. Therefore, at the 5% significance level (p = 0.05, two-tailed), 10 listeners performed well enough to reject the hypothesis that they were selecting the hidden Reference by chance. No listeners identified the hidden Reference few enough times to suspect that an experimental error was the cause of their poor performance in the listening task.

4.2.3 Excluding Listener Data

A listener's data should never be discarded or excluded from analysis without very clear justification. The hidden Reference test results were a reasonable basis on which to recalculate the loudspeaker rankings, using only those listeners who performed at the 5% level.

The binomial percentage test was performed with this reduced data set, remembering that the sample size was now only 30 (10 listeners rating each pair once with each of 3 musical extracts). The smaller sample size meant that a greater critical percentage, 65.5%, was required at the same significance level to confidently say that one loudspeaker design was better than the other. Despite this, the inclusion of only those listeners who demonstrated a reasonable ability to perform the task

resolved one of the ambiguous (non-significant) pairs, as shown in *table 2*. This is enough to discount rankings *4)* and *5)* from *section 4.1*: the ones that were already treated with suspicion as they should have been impossible, excluding an experimental fault.

A/B Pair	d1	d2	d1%	d2%
RC	26	4	86.7	13.3
RD	30	0	100.0	0.0
RE	30	0	100.0	0.0
RF	21	9	70.0	30.0
RG	28	2	93.3	6.7
CD	30	0	100.0	0.0
CE	28	2	93.3	6.7
CF	14	16	46.7	53.3
CG	17	13	56.7	43.3
DE	0	30	0.0	100.0
DF	0	30	0.0	100.0
DG	0	30	0.0	100.0
EF	1	29	3.3	96.7
EG	10	20	33.3	66.7
FG	23	7	76.7	23.3

Table 2: Distribution of subjective votes for each loudspeaker pair, only using data from the top 10 listeners, selected on the basis of a hidden-reference test.

4.3 Investigating Programme Dependence Using Chi-Square

The three extracts of music used to evaluate the loudspeaker designs differed in several ways, such as timbre, meter, and arrangement complexity; however, they were well matched in certain features, primarily their low frequency content. The reason for using different extracts was to look for evidence of programme dependence- the effect discussed by a number of audio researchers (though not always under this name) whereby different characteristics of a loudspeaker become audible with different stimuli. Put simply, listeners' judgement of a loudspeaker may vary depending on the kind of music it is reproducing. *Section 4.3.1* briefly explains the chi-square test (X^2), the method used to compare results across all three extracts, followed by analysis and conclusions in *section 4.1.2*.

4.3.1 The Chi-Square Test For Independence

Also known as Pearson's chi-square, and the chi square test for independence[4], the X^2 ('kigh-square') test looks at differences between categorical variables i.e. the number of counts in each of two or more categories. As such, it performs a similar task to that of the parametric technique, Analysis of Variance (ANOVA).

The 'observed' frequencies occurring in each category are arranged in a special form of table, known as a crosstabulation. For each cell in the table, an 'expected' frequency is also calculated- the number of counts in that category that you would expect to find due to chance alone[15]. The null hypothesis for the test is that the (categorical) variables are not related i.e. that they are independent of each other. As such, the observed and expected frequencies in each cell should be very similar, differing only due to random variations of sampling error. The overall magnitude of the differences between observed and expected frequencies is reflected in the value of the X^2; the larger the differences, the larger the X^2 value. Referring to a table of critical values for the X^2

distribution will show the probability of getting a figure of that magnitude due to sampling error alone; from this, it can be concluded whether differences between the variables are large enough to have been caused by some association between them, or if they are just because of small natural variations between the samples. It is important to note that this test does not tell you anything about the nature of the relationship between the variables or how strong it is, only the likelihood that one exists.

Note: The test statistic in this case is based on the chi-square distribution which only has positive values. The experimenter does not therefore have to choose between a one- or two-tailed test (which was the case for the *z-test* mentioned earlier when using a normal approximation to the binomial distribution).

Snedcor and Cochran[18] (p250) give a good example of how to calculate the value of the chi-square statistic, useful if calculating from a table by hand.

4.3.2 Looking For Differences Between The Extracts

A table (crosstab) was constructed for the number of times each design was chosen for each extract. Brief inspection of *table 3* shows that the observed and expected frequencies are generally very similar; this is an early indication that the value of X^2 for this table should be low.

			Extract			
			DR	DS	SW	**Total**
Design	R	Count	87.0	92.0	95.0	274.0
		Expected	91.3	91.3	91.3	274.0
	C	Count	74.0	75.0	71.0	220.0
		Expected	73.3	73.3	73.3	220.0
	D	Count	6.0	3.0	3.0	12.0
		Expected	4.0	4.0	4.0	12.0
	E	Count	41.0	34.0	36.0	111.0
		Expected	37.0	37.0	37.0	111.0
	F	Count	74.0	84.0	84.0	242.0
		Expected	80.7	80.7	80.7	242.0
	G	Count	63.0	57.0	56.0	176.0
		Expected	58.7	58.7	58.7	176.0
		Total	345.0	345.0	345.0	1035.0

Table 3. Crosstabulation for loudspeaker design vs. musical extract

The value of X^2 was calculated (manually and then verified in SPSS) to be 3.994 (3 s.f.). Referring to a table of critical values for the chi-square distribution, a table with 10 degrees of freedom (*(Rows-1)(Columns-1)*) and a X^2 of 3.994 has a *p* value between 0.950 and 0.900; calculated in SPSS, the exact *p* value = 0.948.

The null hypothesis, H_0, for this test is that the number of times a design was selected is independent of which extract the pair was auditioned with. The alternative hypothesis, H_A, is that design selection was not independent of extract, implying that the selection of a design was influenced by which piece of music it was reproducing. The very high *p* value (0.948) does not lead

to rejection of the null as it suggests that there is approximately a 95-in-100 chance that the differences in frequencies between extracts are due only to random sampling variations.

4.4 Correlation Between Subjective And Objective Rankings

The subjective data here is the rankings of the loudspeaker designs based on listener judgements of reproduction accuracy. Similarly, the objective data is the rankings of the same designs from analysis with an algorithm that processes their impulse response. As mentioned in *section 3.1.1.,* this algorithm aims to judge the accuracy of the bass reproduction of each loudspeaker by taking into account its time and frequency behaviour. This is based on the Modulation Transfer Function (MTF); for more details of the technique and signal processing, the reader is referred to earlier papers by the same authors[19,20]. It should be noted that the subjective (listening test) ranks and objective (MTF) ranks are completely separate data sets, only linked by the fact that they both have the same loudspeaker designs to compare in their different ways. It must also be mentioned that The MTF algorithm has a number of parameters that can be modified. The values given here are from the algorithm as it was at the time of testing. In the future it is anticipated that the parameters, and therefore the resulting MTF scores, will change, though the loudspeaker responses themselves will not be altered. Specifically, a weighting is being developed that may help the MTF scores reflect subjective impression more accurately.

A test was performed to find the strength of correlation between the subjective and objective rankings. The technique is introduced in *section 4.4.1*, and the results presented in *section 4.4.2*.

4.4.1 Correlation Methods And Significance

Spearman's rank-order correlation coefficient (Spearman's rho, r_s) was used to investigate the strength of relationship between the loudspeaker rankings. This is a simple calculation based on the difference between corresponding rank values in the two data sets. A *t-test* was performed with each r_s value to find its significance i.e. the probability that this strength of correlation was due to chance alone. (This use of a *t-test* is clearly demonstrated in Field[15], p366.)

Spearman's rho indicates the linear relationship between two variables, and like other measures of correlation, great care must be taken when interpreting the results. It must always be remembered that correlation does not imply causation, i.e. just because two variables appear to be strongly linearly related, one does not necessarily cause the other. Apart from the possibility that the apparent association may be due to sampling error, the real cause of the evident effect might be, and often is, an unknown third variable. It can also be helpful to plot the variables against each other in a scatter plot before carrying out any formal analysis; linear trends will be at least partially visible. Finally, correlation coefficients, where apparently significant, must be considered within the *practical* context of the experiment; common sense and restraint must be used when drawing conclusions from the analysis.

4.4.2 Magnitude Of The Correlation

The five possible subjective rankings from *section 4.1* were compared with two objective rankings: one by the MTF algorithm score, and the other by frequency response alone (bass extension). The Spearman's rank correlation coefficient, r_s, was calculated for each of these using *equation 4*:

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} \qquad (4)$$

Where *D = difference in rank between the two data sets*, *n = sample size* (here *n = 6*)

Note: The working is not shown for brevity, but Field[15](p180) demonstrates the technique very clearly.

The results of the correlation analysis are shown in the results are shown in *table 4.* In this case, the null hypothesis, is that there is no correlation between the subjective and objective rankings i.e. $\rho = 0$. The alternative hypothesis is that there is a correlation i.e. $\rho \neq 0$. Note that this is a non-directional alternative hypothesis**,** so the test should be two-tailed.

Possible listening test rank	Rank by FR	Rank in MTF, low to high
R F C G E D	1.000**	.714
R F G C E D	.943**	.543
R C F G E D	.943**	.829*
F R C G E D	.943**	.543
F R G C E D	.886**	.371

**. Correlation is significant at the 0.01 level (2-tailed)

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4. Possible rankings from listening tests and the associated Spearman's rank correlation coefficient based on frequency response extension and MTF score.

The values of r_s are clearly higher for the frequency response (FR) comparisons than the MTF ones. SPSS automatically tests the significance of the correlation coefficient, as indicated in the table using asterisks; if computing the coefficients manually, a *t-score* for the sample can be calculated using *equations 5 and 6*; these can then be looked up in a table of critical values for the *t*-distribution to find the significance values.

$$s_r = \sqrt{\frac{1-r^2}{n-2}} \quad (5) \qquad \text{and} \qquad t_{sample} = \frac{r_s - \rho}{s_r} \quad (6)$$

Where *n = sample size,* r_s *= Spearman's rank correlation coefficient*, *rho,* ρ *= hypothesised population correlation coefficient,* and s_r *= standard error*[4]*.*

Given the results of this correlation analysis, the original loudspeaker designs were compared together to see if this outcome should have been expected. It could be seen when plotting all 6 of the model responses together that, certainly in the clearly audible region above 40Hz, the difference between designs was purely a difference in low frequency extension (bass output). In this context, the strong correlation between frequency response ranking and subjective ranking made sense.

4.5 Summary Of The Analysis

The experimental loudspeaker designs were ranked according to the listening test count data. A normal approximation to the binomial distribution was used to decide which subjective differences between loudspeakers in each pair were significant. Based on the results form 23 listeners, 5 different rankings of the 6 experimental designs were possible.

The exact binomial distribution was used to assess each listener's individual performance at identifying a hidden Reference design. No listeners performed poorly enough to suspect an experimental flaw, based on the outcome of a two-tailed significance test at the 5% level; it was found that 10 out of 23 listeners performed at the specified significance level. The loudspeaker ranks were then recalculated with just the data from those 10 listeners; the disagreement between

responses for each pair of designs was found to be reduced in nearly all cases, and the number of possible rankings decreased to 3.

The chi-square test was used to analyse the distribution of results between musical extracts; it was concluded that the loudspeaker designs were selected independently of the piece of music being used to evaluate them. There was therefore no evidence of programme dependence.

Finally, the strength of the relationship between the loudspeaker rankings was investigated; subjective bass reproduction accuracy was compared to two objective rankings, one based on frequency response (bass extension), and the other on a Modulation Transfer Function analysis (time and frequency behaviour). Spearman's rank-order correlation coefficient was calculated for each of the possible subjective rankings with both objective rankings. It was clear that the listeners' judgement of bass reproduction accuracy correlated very strongly with the frequency response extension of the loudspeaker designs. When the experimental loudspeaker responses were all plotted together it became clear that the only real difference between them, certainly in the audible region, was in their relative bass extension; thus, the results of the correlation analysis supported the acoustic evidence.

5 SUMMARY

After a short review of some useful introductory statistical sources, an ABX listening test case study was presented. Important features of the experimental design and execution were described before an explanation of how the data was analysed. An introduction to each statistical technique was presented, followed by the results and conclusions from the listening tests. Particular attention was paid to non-parametric statistical analysis throughout the paper; this was appropriate for the kind of data produced by *two-alternative forced choice* tests, but is not often addressed fully in the audio literature. Analysis methods used in the case study included binomial percentage calculations when approximating the normal distribution, the chi-square test for independence, and Spearman's rank order correlation coefficient.

Acknowledgements: Thanks to Will Evans at the University of Surrey for correspondence about statistical sources when writing this paper.

6 REFERENCES

1. Lipshitz, S.P. and J. Vanderkooy, The Great Debate: Subjective Evaluation. Journal of the Audio Engineering Society, 1981. 29(7-8): p. 482-91.
2. Geddes, E.R., L.W. Lee, and R. Magalotti, Subjective Testing Of Compression Drivers. Journal of the Audio Engineering Society, 2005. 53(12): p. 1152-1157.
3. EBU, Tech. 3286-E: Assessment Methods For The Subjective Evaluation Of The Quality Of Sound Programme Material – Music. 1997, European Broadcasting Union
4. Argyrous, G., Statistics For Research : With A Guide To SPSS 2nd ed. 2005, London: SAGE Publications.
5. ITU-R, Recommendation BS.1116-1 Methods For The Subjective Assessment Of Small Impairments In Audio Systems Including Multichannel Sound Systems. International Telecommunications Union, 1994.
6. Bech, S. Listening Tests on Loudspeakers: A Discussion of Experimental Procedures and Evaluation of the Response Data. In AES 8th International Conference: The Sound of Audio 1990: Audio Engineering Society.
7. Bech, S. and N. Zacharov, Perceptual Audio Evaluation : Theory, Method And Application 2006, Chichester John Wiley.
8. Leventhal, L., Type 1 And Type 2 Errors In The Statistical Analysis Of Listening Tests. Journal of the Audio Engineering Society, 1986. 34(6): p. 437-453.

9. Burstein, H., Approximation Formulas For Error Risk And Sample Size In ABX Testing. Journal of the Audio Engineering Society, 1988. 36(11): p. 879-883.
10. Leventhal, L. and C.L. Huynh, Analyzing Listening Tests With The Directional Two-Tailed Test. Journal of the Audio Engineering Society, 1996. 44(10): p. 850-863.
11. Freund, J.E., Modern Elementary Statistics 12th ed. 2007, New Jersey Pearson Prentice Hall.
12. Meddis, R., Statistics Using Ranks : A Unified Approach. 1984, Oxford: Basil Blackwell.
13. Siegel, S., Nonparametric Statistics For The Behavioral Sciences 1988, New York: McGraw-Hill.
14. Rowntree, D., Statistics Without Tears : A Primer For Non-Mathematicians 1981, Harmondsworth: Penguin.
15. Field, A., Discovering Statistics Using SPSS : (And Sex, Drugs And Rock'n'roll) 2nd ed. 2005, London: SAGE Publications. .
16. Martin, G. Introduction To Sound Recording: Two-Alternative Forced Choice. 2006 (Accessed 03/11/09); Available at http://www.tonmeister.ca/main/textbook/node361.html.
17. McCormick, E.J. and J.A. Bachus, Paired Comparison Ratings. I. The Effect Of Ratings Of Reductions In The Number Of Pairs. Journal of Applied Psychology. 1952. Vol 36(2) p. 123-127.
18. Snedcor, G.W. and W.G. Cochran, Statistical Methods. 6th ed. 1967, Ames: Iowa State University Press.
19. Harris, L.E., K.R. Holland, and P.R. Newell, Subjective Assessment Of The Modulation Transfer Function As A Means For Quantifying Low-Frequency Sound Quality. Proceedings of the Institute of Acoustics, 2006. 28(8): p. 195-203.
20. Harris, L.E. and K.R. Holland, Evaluating Loudspeaker Quality At Low Frequencies: Optimisation Of A Music-Focussed Modulation Transfer Function Technique Proceedings of the Institute of Acoustics, 2008. 30(6).