Audio Engineering Society

Convention Paper

Presented at the 127th Convention
2009 October 9–12 New York, NY, USA

Statistical Analysis of ABX Results Using Signal Detection Theory

Jon Boley[1] and Michael Lester[2,1]

[1] LSB Audio, Lafayette, IN 47905, USA

[2] Shure Incorporated, Niles, IL 60714, USA

ABSTRACT

ABX tests have been around for decades and provide a simple, intuitive means to determine if there is an audible difference between two audio signals. Unfortunately, however, the results of proper statistical analyses are rarely published along with the results of the ABX test. The interpretation of the results may critically depend on a proper statistical analysis. In this paper, a very successful analysis method known as signal detection theory is presented in a way that is easy to apply to ABX tests. This method is contrasted with other statistical techniques to demonstrate the benefits of this approach.

1. INTRODUCTION

Within the audio engineering community, the ABX methodology[1] has become the standard psychoacoustic test for determining if an audible difference exists between two signals. In this experimental procedure, three stimuli are presented. Stimulus "A" is one sound, stimulus "B" is known to be quantitatively different in some way, and the task of the listener is to identify whether stimulus "X" is the same as "A" (i.e. X=A) or the same as "B" (X=B).

If there is no audible difference between the two signals, the listener's responses should be binomially distributed such that the probability of replying "X=A" is equal to the probability of replying "X=B". If reported as *percent correct*, a score of 50% is often interpreted as indicating no perceptual difference between A and B. Although *percent correct* is a very common reporting method, two questions may come to mind: how high does the percent correct need to be to indicate a perceptual difference? and how confident can we be in our interpretation of the results?

We will first review some experimental design considerations, followed by a discussion of statistical analyses, then close with some notes on appropriate ways to interpret and report results from an ABX listening test.

2. EXPERIMENTAL DESIGN

The design of the experiment is the most crucial factor when preparing to perform a listening test. It can be quite frustrating to spend the time and effort of conducting a listening test, only to figure out that you cannot make any reasonable conclusions. In such a case, you would have to fix the problem and start over.

It should first be mentioned that there are quantitative methods for measuring differences between audio sources. If an objective calculation will answer the question being posed, it will often be less costly and/or time consuming than a listening test. However, listening tests are often used to measure the perceived differences of known objective differences. Note also that it is not necessary to perform listening tests on material that have no objective differences.

The first step in any listening test is to clearly define the question you are trying to answer and an associated hypothesis to be tested. It is very important to be specific as possible, while encapsulating all the necessary components. Often, the question will have multiple components and needs to be broken down into smaller questions that can each be answered by a particular test methodology. For example, ABX can only tell you if there is a perceptible difference between two sounds. It cannot tell you that there is no difference between the sounds.

The next step, and perhaps one of the most overlooked, is to identify outside factors or variables that could lead to an incorrect conclusion. The listening test should take place in a suitable room for the particular sounds you are interested in. The loudspeakers should also be appropriate for your test (headphones may be preferable for some studies). You will also want to decide on a comfortable level to present the sounds out. The ITU has standardized many aspects of a suitable listening environment, and standards such as ITU-R BS.1116 are useful as a guide. If you have trouble controlling a variable you are not interested in studying, good practice is to randomize it as much as possible. (Computer programs are quite useful here, because what we perceive as random is often not.)

It is important to keep all environmental variables consistent for all listeners. If something is allowed to change (e.g., room configuration, loudness, etc), it can be very difficult to interpret the results. In fact, any perceptible difference may have been caused by one of these variables that you did not control for.

The audio material (music, speech, etc) should be representative of typical material for the system being evaluated, but it should also be critical material. In other words, it should push the system to its limits and bring out any potentially audible differences. If a high frequency phenomenon is being tested, telephone quality speech may not be an acceptable stimulus. Also consider how much and what variety of listening material is necessary to prove your hypothesis. Many experiments choose varied styles of music as stimuli, but each experiment should have different stimuli based on the experimental question and what is necessary to prove your hypothesis. For instance, if a device is known to have a quantitative difference in time response qualities such as transient attacks, files should be included with impulsive material such as isolated drum hits or castanets.

For many studies, the experimental subjects should be qualified expert listeners. Expert listeners are those who have been specifically selected for their listening skills- they are typically quite sensitive to the specific audio qualities you are interested in and they are very consistent in their assessments.[2]

The experiment should be double-blind if at all possible- the test administrator should not know the answers. This will ensure that the administrator does not accidentally give subtle cues to the listener.

The order of stimulus presentation should be randomized. It is quite possible for the presentation order to affect the listener judgments, and just like any other variable, this should be controlled for.

It is generally useful to ask listeners to write down their impressions & opinions after the experiment. Often, these comments will reveal aspects that you as the experimenter never considered.

In fact, that leads us to our next point- it is good practice to always run a small pilot test before beginning the actual experiment. This will allow you to collect a few comments, test the setup, validate the experimental design, and verify that your statistical analysis of the results can lead to useful conclusions. If you are using expert listeners, they could be helpful here. If your specific test requires naïve listeners (very rare for an ABX test), a separate panel of listeners should be used for the pilot and the full test.

The listeners should all receive the same instructions (again to avoid unexpected variability).

A written instruction sheet works well, as do audio/video recordings.

The experimental design considerations of off-line experiments (e.g., listening to recorded audio files) are often different from those for real-time or on-line experiments – those that require running audio through a device under test (DUT) during a listening test experiment. One must understand the ramifications of isolating the intended test phenomenon from the tangential equipment such as cables, gain stages, loudspeakers, etc. For example, if an experiment attempts to study the audible difference in dynamic range of two digital formats such as 16-bit and 24-bit, one must consider the effects of the devices used to conduct the test. Unrelated signal chain considerations are necessary such as analog circuitry, routing, shielding, converter quality, and so on. Consider that if one DUT has significant advantages or disadvantages due to an unrelated variable, the results may not be indicative of the intended test purpose.

Real-time experiments will require sophisticated switching mechanisms in order to eliminate the possibility that a subject can hear a switching artifact which would give away the answer to a test such as ABX. Real-time experiments can be difficult to design as double-blind studies due to an increased need for special hardware to create such a test; however double-blind testing is important and should be considered with care. Both real-time and on-line experiments are particularly sensitive to amplitude and delay matching.

Slight variations in either amplitude or phase can again reduce the validity of the experiment. Attempts to correct audio characteristics often require additional equipment in the signal chain. In this case it is crucial to include duplicate equipment in all signal paths. It is not acceptable to place a gain section in one path and not the other to match gains. In fact, it is prudent to split the gain difference and apply some gain or attenuation to both paths rather than altering just one path.

Other considerations are intuitively obvious, but are sometimes overlooked and are worth mentioning. For example, it is important to consider the experimental purpose when choosing listening audience. If the listening test involves high frequency response, those with limited high-frequency hearing may not be suitable subjects.

Proper experimental design is paramount and is the foundation for conclusive statistical analysis. Fully describing proper experimental design is outside the scope of this paper, but an excellent reference on experimental design can be found in Bech and Zacharov's book[3].

3. STATISTICAL ANALYSIS

Proper statistical analysis is an important step for any scientific inquiry, and is certainly important in psychoacoustics. For a discrimination task like ABX, there are two primary analysis techniques, as described in the next sections. When considering the two analyses, remember that the ABX test is a Bernoulli trial (a test in which the outcome can be classified in one of two mutually exclusive and exhaustive ways such as success or failure) and refer to the following table of response pairs.

Stimulus/Response	Response "A"	Response "B"
X = A	Correct	Incorrect
X = B	Incorrect	Correct

Table 1. Simple response matrix

3.1. Binomial Distribution

One option is to use a simple application of the Binomial Distribution[4] to determine the statistical significance of test results. Two assumptions are necessary to employ the binomial distribution. The first is that the ABX testing program (software or otherwise) is randomly distributing correct answers of X=A and X=B throughout the test. The second assumption is that when a subject is unable to identify the correct answer, the response is random and uncorrelated to the audio tests.

Given a Binomial Distribution, a specific number of trials, and an observed number of correct identifications, a *p-value* can be calculated (or looked up in a table). This value represents the probability of randomly getting more correct identifications with the same conditions. For example, given 100 trials and 60 correct answers (and assuming that the subject has a 50% chance of randomly guessing the correct answer), the probability of randomly getting at least 60 correct answers is 0.0284. In other words, there is only a 2.84% chance that the subject was just randomly guessing (although the reader should also recognize that this implies that approximately 1 in every 35 experiments will randomly produce 60 or more correct answers).

Let us explore the derivation of the binomial distribution and its application to ABX tests in order to better understand its application. This analysis considers the total number of successes rather than the order in which they occurred. A random variable 'S' equals the number of observed

successes in *n* Bernoulli trials, the possible values of S are 0, 1, 2...,*n*. If *s* successes occur where *s* = 0,1,2... ,*n* then *n*-*s* failures occur. The number of ways of selecting *s* positions for the *s* successes in the *n* trials is noted as the following:

$$\binom{n}{s} = \frac{n!}{s!\,(n-s)!}$$

Since the trials are independent and since the probabilities of success and failure on each trial are p and q = 1-p respectively, the probability of each of these ways is

$$p^s * (1-p)^{n-s}$$

The Probability Mass Function, f(*s*), is the sum of the probabilities of these $\binom{n}{s}$ mutually exclusive events

$$f(s) = \binom{n}{s} * p^s * (1-p)^{n-s}$$

These probabilities are known as the binomial probabilities, and the random variable 'S' is said to have a binomial distribution[5].

A table of binomial distributions can be found at http://www.lsbaudio.com/ABX/index.html

We can calculate these binomial probabilities for each *s* out of n possible outcomes and display them as a distribution such as the following example of 10 trials (see Fig. 1).

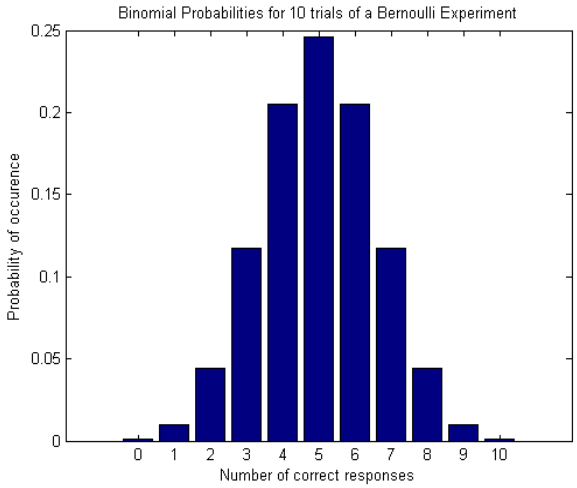

Figure 1. Binomial probability mass function (n=10)

In the analysis of an ABX experiment, the test for a significant result is that there is a very low probability that the resulting *percent correct* was due to chance – that the subject guessed at the responses and happened to get a particular *percent correct* score. We can use the knowledge of the binomial probability to calculate the inverse cumulative probability of getting *s* or more correct answers in a given random trial. The inverse probability is used because in any give experiment there is a 100% chance of getting 0 or more correct identifications.

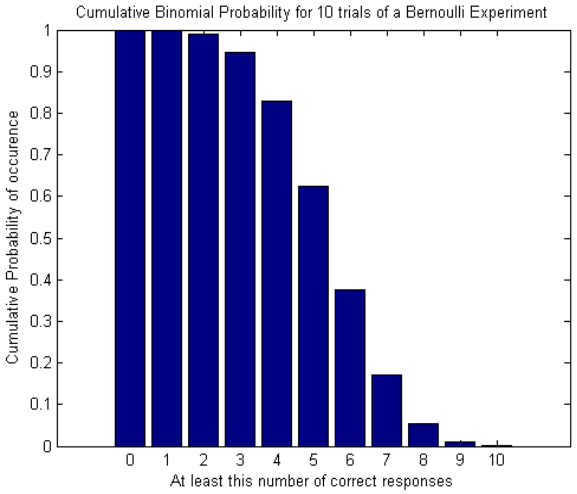

Figure 2. Inverse binomial cumulative distribution function (n=10)

Fig. 2 shows that there is a particular chance involved in randomly getting *s* responses correct out of n trials. To be more confident that the result is not due to chance, more correct responses are needed. As an example, if there were 8 out of 10 correct responses from a Bernoulli experiment, there is a 5.5% chance that it was due to random guesses. The reader should also recognize that given random responses for 100 tests, 5 test results would show at least 8 out of 10 correct responses.

It is also important to consider the reversal effect. For example, consider a subject getting 1 out of 10 responses correct. According to the cumulative probability, it may be tempting to say that this outcome is very likely: 99.9% in this case. However, consider that your test subject has been able to incorrectly identify 'X' 9 out of 10 times. The cumulative probability shows the likelihood of getting at least 1 out of 10 correct responses. The binomial probability graph shows that to get exactly 1 out of 10 correct responses has a probability of 0.97%. It is possible that the subject has reversed his/her decision criteria. Rather than identifying X=A as X=A, the subject has consistently identified X=A as X=B. The difference in the stimuli was

audible regardless of the subjects possible misunderstanding of the ABX directions or labeling.

Typically, a 95% confidence level is sufficient for psychoacoustic experiments. In other words, we want to be confident that less than 5% of completely random test results would indicate a perceptible difference. The number of correct responses necessary to obtain a 95% confidence level are shown below:

Trials:	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
≥	8	8	9	9	10	11	11	12	12	13	14	14	15	15	16	17
or ≤	1	2	2	3	3	3	4	4	5	5	5	6	6	7	7	7

Table 2. Results required for a 95% confidence level

For example, given 20 trials, a result of ≥ 14 correct responses, or ≤ 5, indicates a significant perceptual difference between A and B.

Let us now consider the possibility that the assumptions mentioned at the beginning of this section are invalid. The binomial distribution statistical analysis of the ABX tests is best used for large number of trials. Unfortunately, it is difficult to design a successful experiment in which your subjects will not get fatigued after listening to the same audio file a large number of times. Consider the differences in probability and statistical distribution. Small numbers of trials may have a 50% probability of the correct answer being X=A and X=B, while the actual statistical distribution of X=A and X=B may not be even. As shown above, a random number generator may produce eight X=A trials and two X=B five times out of every 100 instantiations of the algorithm.

If the second assumption is not true, the subject will not be a true random number generator when responding to audio files that are indiscernible. Violating this assumption compounds the effects of violating the previous assumption. As a simple example, if the correct responses are unevenly distributed and the subject were to be biased toward consistently answering either A or B, similar to how some subjects respond if they are unsure of the answer, a contrived statistical significance may emerge.

The next section presents a more robust statistical analysis of the ABX test. More examples of unintended false conclusions can be found at http://www.lsbaudio.com/ABX/index.html

3.2. Signal Detection Theory

In the above analysis, we have assumed that the listener is equally likely to incorrectly answer "X=A" when X=B and "X=B" when X=A. If we instead consider that any given listener may be more or less conservative about his/her decisions[6], the statistical analysis must change. In this scenario, we must separate sensitivity (ability to discriminate between A and B) from bias (e.g., conservative/liberal responses and/or bias due to the experimental setup). To pursue a more detailed analysis, we must consider that there are in fact four possible stimulus/response pairs:

Stimulus	Response	Example Value
1) X=A	X=A	30
2) X=A	X=B	20
3) X=B	X=A	10
4) X=B	X=B	40

(The example values represent the number of responses for each case, out of a total of 100 trials.)

Note that, in an unbiased experiment, we expect cases 1 and 4 to be equally probable and cases 2 and 3 to be equally probable. If the results suggest that this is true (i.e. the proportion of cases 1 and 4 are equal, and the proportion of cases 2 and 3 are equal), it is sufficient to report the fact that the results were unbiased along with the *percent correct* and the statistical significance based on the binomial distribution. If, however, the proportions suggest a biased result (as illustrated in Fig. 3), signal detection theory may be used for further analysis.

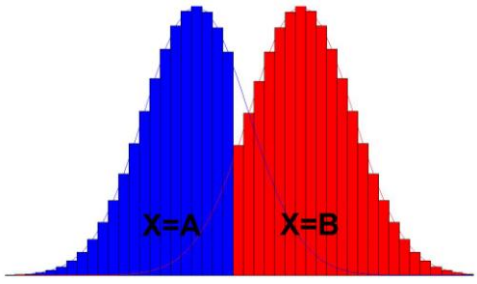

Figure 3. Two binomial distributions with a slight bias toward answering "X=B"

In detection theory analysis, we consider the hit rate, *H*, and the false alarm rate, *F*, of either A or B. Using the example values given above,

$H_A = 30/50 = 0.6$ (case 1)

H_B = 40/50 = 0.8 (case 4)
F_A = 10/50 = 0.2 (case 3)
F_B = 20/50 = 0.1 (case 2)

If we assume a Gaussian distribution, we can convert these scores to a standard normal representation such that

$z(H_A) = 0.2533$
$z(H_B) = 0.8416$
$z(F_A) = -0.8416$
$z(F_B) = -0.2533$

where z(·) represents the z-score, or the point at which the standard normal distribution is equal to the value specified.

The metric *d'* ("dee-prime") can then be looked up in a table like that found in Table A5.3 of MacMillan and Creelman's book[7]. For our example, we can calculate

$$z(H_A) - z(F_A) = z(H_B) - z(F_B) = 1.095,$$

and look in the table under "Differencing" and find a *d'* value of 1.765. (Differencing is likely the best model for ABX, as argued by Hautus and Meng[8]. Basically, it models the ABX task as a minimization of the difference between A and X, then B and X.) The bias, or *c* (for *criterion*), can be calculated as

$$c = \frac{z(H_A)+z(F_A)}{-2} = \frac{z(H_B)+z(F_B)}{2} \quad (1)$$

For our example, *d'*=1.765 and *c*=0.2941.

The variance of d' can be calculated using the equation

$$var(\mathrm{d}') = \left(\frac{\mathrm{H}(1-\mathrm{H})}{\left(\frac{\mathrm{N}}{2}\right)\mathrm{z}^2(\mathrm{H})}\right) + \left(\frac{\mathrm{F}(1-\mathrm{F})}{\left(\frac{\mathrm{N}}{2}\right)\mathrm{z}^2(\mathrm{F})}\right) \quad (2)$$

where N is the total number of trials. In our example, the variance is 0.0793, so the standard deviation (square root of the variance) is 0.2816. Now, if we want to calculate a 95% confidence interval for *d'*, we simply calculate 1.96 standard deviations and say that the 95% confidence interval for *d'* in our example is 1.765±0.5519 (or the range from 1.1021 to 2.3169).

The application of signal detection theory requires knowledge of the subject's response and the correct answer for every trial of a listening test. Note that many ABX software packages available as of the publish date of this paper do not provide the experiment designer this information.

4. REPORTING & INTERPRETING RESULTS

This section gives some suggestions of what to report (and what not to report) when describing the results of a listening test. It may not be necessary to include all of the suggestions; however, if the information is available it should be included in the results in some form.

The following are suggestions to include for all listening tests:

- The hypothesis should be defined and related to the experimental design method selected (ABX)
  - What is your criteria for proof? How many people need to hear a difference? On how many different stimuli?
  - Was the hypothesis proven?
- A method section that considers all practical aspects related to the experiment. Sufficient detail should be present such that the experiment could be reproduced or repeated.
- Details of any apparatuses used to conduct the experiment
- Describe the listening environment, test setup, and listening material
- Was the test double blind?
- Were the assignments to A/B/X random?
  - Describe the actual distribution of correct answers?
- Number of subjects
- How were the subjects chosen?
- Number of trials per subject
- If you found a perceptible difference, report the confidence level.
- Report any individual listeners who appeared to hear a difference (along with the statistics for that listener).
- Were there any correlations to the particular subjects who could hear differences?
- Relevant comments recorded on subject comment sheets
- Be willing to present the raw data using websites or email, but be sure to remove any personally identifiable information.
- Provide the objective differences in the results to qualify the perceived differences.
- Present any and all special experimental design considerations. Explain how variables were controlled.
- Comparisons to related works
- References providing background motivation for the experiment

For a binomial analysis, it is important to demonstrate that the results were unbiased (overall, and for each subject). For any significant results (e.g. >95% confidence level), the percent correct should be reported along with confidence level, number correct and total number of trials. A statistically significant result indicates that there is a low probability that the listeners were merely guessing. (However, the reader should understand that it is completely possible to randomly obtain statistically significant results.) In the case of an ABX test with results at the 95% confidence level, we can say that there is only a 5% chance that the listeners were randomly guessing. Typically, we would be comfortable claiming that there is probably an audible difference between the two signals.

For a detection theory analysis, the important statistic to report is the calculated value of *d'* and the associated confidence interval (although it may also be beneficial to report the bias). In our example, the confidence interval for *d'* was 1.765±0.5519, In signal detection theory, a *d'* value of zero indicates no perceptible difference. Because the 95% confidence interval for *d'* does not include zero, we can be comfortable saying that the listener was probably able to discriminate between the two signals.

Here are some things to be sure not to do:

- Do not lump people together when doing ABX test analysis. Each subject introduces a subjective criteria variable that cannot be combined with others. Instead, consider reporting the number of people who detected a difference.
- Do not conclude there is no difference. ABX can only prove differences in audio files, not prove that there is no difference.
- Do not lump stimuli results together unless they are quantitatively similar. When lumping sample sets together, variables can be combined that should not be combined.

5. CONCLUSIONS

More sophisticated analysis such as signal detection theory may help prevent some potential false conclusions by accounting for biases in the numerical results. However, note that the foundation of a good experiment is the experimental design. The statistical analysis cannot make up for poor experimental design.

The binomial distribution method of ABX analysis is an appropriate analysis but is more likely to produce accurate results only when larger sample sets are used. It is possible to report audibility thresholds with lower percent correct scores using signal detection theory. Due to the larger number of samples needed for proper binomial distribution analysis, the signal detection theory may be desirable in that it may be possible to prove the hypothesis with fewer total number of trials.

6. REFERENCES

[1] Munson, W.A. and M.D. Gardner. Loudness patterns - A new approach. Journal of the Acoustical Society of America, 22:177–190, 1950.

[2] ISO 8286-2:2008: Sensory analysis -- General guidance for the selection, training and monitoring of assessors – Part 2: Expert sensory assessors. International Organization for Standardization, Geneva, Switzerland

[3] Bech, S., N. Zacharov. Perceptual Audio Evaluation – Theory, Method and Application. First Edition. West Sussex: John Wiley and Sons, Ltd. 2006.

[4] Clark, D. High-Resolution Subjective Testing Using a Double-Blind Comparator. JAES Volume 30 Issue 5 pp. 330-338; May 1982

[5] Hogg, Robert and Tanis, Elliot. Probability and Statistical Inference. Sixth Edition. New York: Prentice Hall College Division. 2001.

[6] Harris, J.D., Remarks on the Determination of a Differential Threshold by the So-Called ABX Technique. J. Acoust. Soc. Am. 24, 417 (1952).

[7] MacMillan, N.A., C.D. Creelman. Detection Theory, A User's Guide. Second Edition. MahWah, New Jersey: Erlbaum. 2005.

[8] Hautus, M. J. and X. Meng (2001). Decision strategies in the ABX (matching-to-sample) psychophysical task. Perception & Psychophysics, 64, 89-106.