

Design of a Listening Test

Laura Enflo

Statistical Methods, GSLT

May 17, 2009

Listening tests are frequently used for detecting, or stating a lack of, differences between given samples. In addition, a common wish is to get results with statistical significance, which means how unlikely it is that the outcome occurred by chance. Another wish, when having statistical significance at some level, is to get a high power number. The power value is a measure on how sure the tester can be of the result – the higher number, the better.

This investigation deals with a voice source which will be adapted to Matlab and two (or more) parameters which will be changed in order to see their impact on voice quality. The parameters will be chosen according to the results from the new voice source program and studies of previous research. A listening test will be made after this procedure and the task for the Statistical Methods course is to decide how to form this test and which statistical analysis methods that can and should be used. This listening test will consist of at least two parts; one for each parameter that is changed. The listener will hear sound sample pairs: one sample that is changed and one which is left unchanged. This means that the given answer for each pair is either correct or wrong and the probability for choosing each sample of the two, if no difference is there, would obviously be approximately 50 %.

Dependence versus independence

In this listening test, sample pairs originating from the same file will be compared, with only one parameter change to make a difference between them. These pairs would be considered dependent, like all variables measured or registered. Several significance tests, for example the chi-square test, are working for independent variables only. The term independent are referring to variables that are manipulated in some way, for example the distances which, lets say, ten athletes can run in five minutes. We take to heart that this listening test is based on dependent variables and therefore limit the investigation to significance tests designed for those.

Dependent, paired T-test

A dependent, paired T-test means that two correlated samples at a time are compared with each other and that the sample groups are supposed to have equal, but unknown, variance and be normally distributed (degrees of freedom = $n-2$) if n is not too small, which means that the t-test is the same as the z-test in this case. The number of samples in each of the two groups must be equal as well.

If we just want to know if there is a significant difference between the sound samples of type A versus type B, a two-tailed test is an adequate tool. On the other hand, if we just want to count the correct answers (given, for example, that we should identify a difference in decibel between the samples) a one-tail test is enough and would also generate a higher significance level than the two-tailed test. In short: when we want to

show that there is a difference between samples A and B, a two-tailed test is sufficient. If we just want to show that A is always greater than B (not that B is greater than A), a one-tail test is correct to use. Hence, a two-tailed test is preferable in this listening test and can easily be carried out with, for example, the computer program SPSS.

Analysis of variance

ANOVA – analysis of variance – is a collection name for several statistical methods aiming to compare the means for several different groups and see if they are equal. In this listening test we only use two groups, which mean that the ANOVA test is the same as the two-sample t-test. Therefore, we have no reason to consider this.

P-value

The outcome of the significance test is given as p equalizing a certain number. P stands for probability for the null hypothesis to be true. If $p=0.05$, we have a 95 % chance that the, for example, detected difference between two groups is indeed there. A probability $p=0.01$ gives a 99 % chance in the same way.

Number of listeners & Svante's Test Design

One major question when drawing up a listening test is how many listeners that is needed in order to have a higher probability for statistical significance to be reached as the random errors are marginalized.

One method used in the electroacoustics course at KTH would after a modification to becoming a two-tailed test be suitable for this listening test. The test made in the laboratory experiment aims to find out the threshold for detecting differences (a lowering) in dB for music on CD. Two samples are played over headphones: one changed and one unchanged. The listener presses a button when the correct sample is given. This means that the answers can be either right or wrong. A listening test consisting of three sample pairs played in a row, gives us 2^3 possible sequences. In this test, only the row with three correct answers will be counted on as a positive outcome, which give us $1/8=12.5$ % chance of a false positive answer. (A two-tailed test would also consider the all-wrong sequence). This means that $7/8=87.5$ % is the level of significance. This number is not very promising (see the P section) and we realize that more sample pairs should be played for the listener. Therefore, one could decide to play five sample pairs in a row instead. This means $2^5=32$ possible outcomes of which, if we would allow one fault, and six sequences would be considered correct. The level of significance is now $26/32=81.25$ %. Comparing this result to the previous one with three sample pairs, we realize that we have to decide in advance how many sample pairs we will listen to, or else the level of significance will be lowered. Acceptance of one wrong answer means that one more sequence would be considered correct for the three sample pairs; hence, the level of significance is lowered to $25/32=78.125$ %. For the one-tail test, seven sample pairs need to be played in a row for the significance level to reach 99 %. For the wanted two-tailed test, we would have to listen to 20 sample pairs before the significance level could reach 90 % and 40 sample pairs for 95 % significance. A 99 %

significance level would not be realistic in this case, since the human ear gets tired long before that is achieved.

Summary

After considering different options for the listening test, it seems reasonable to let the subjects listen to at least 40 sample pairs for each of the two parameter experiments (with a pause between them). After collecting the data and especially if subjects do not have 100 % correct answers, a dependent, paired t-test will be performed in SPSS.

References

Granqvist, Svante (2008) *Elektroakustik, laboration B2, lyssningstest*. URL: <http://www.csc.kth.se/utbildning/kth/kurser/DT2400/ablab.pdf> [visited May 17, 2009]
Stockburger, David W. (1998) *Introductory Statistics: Concepts, Models, and Applications*, WWW Version 1.0.