Committee T1Y1.1
Contribution

Document Number: T1Y1.1/89-553

Standards Project: Digital Encoding of System M-NTSC Television Signals for Broadcast Quality Transmission at the DS3 Rate.

Title: Statistical Design and Analysis of Test Program for Selecting Codec Standard for Broadcast Quality NTSC Television at DS3.

Source: National Telecommunications and Information Administration
Institute for Telecommunication Sciences
ITS.N3, Boulder, CO 80303
Edwin L. Crow

Date: October 2, 1989

This document comments on and supplies two alternative statistical methods for designing and analyzing the test program for selecting a standard codec for video transmission that is described in T1Y1.1 Document Number T1Y1.1/89-060R2.

Any suggestions or comments on this document can be addressed to:

Ralph C. Brainard
AT&T Bell Labs
Room 4C534
Holmdel, NJ 07733
201-949-4147
FAX 201-949-3697

Statistical Design and Analysis of Test Program for Selecting Codec Standard for Broadcast Quality NTSC Television at the DS3 Rate.

1. Introduction

Document Number T1Y1.1/89-060R2 proposes a procedure for testing proponent codecs for transmission of NTSC television at the DS3 rate and selecting one of them as the standard. The purpose of the present document is to describe design considerations that will enable the statistical significance of the mean subjective responses to be stated, thus determining whether there is a clear winner. This purpose of the test procedure is described in Section 7 of the referenced document, which is assumed to be available and details of which will not be repeated except as needed.

It is assumed that the objective tests will be used only to assure that all the candidate codecs meet certain minimal requirements or in the case the subjective testing does not result in a clear winner. Hence, the objective tests are not considered further in this document.

It is assumed that several codecs will be tested by using them on tapes of several scenes subjected to several different tests (basic quality, error performance, recovery from gross disturbances, chromakey, slow motion, tandem operation) viewed by many observers at three or more locations. It is assumed that the observer response records will be analyzed by two independent laboratories.

In order that statements of statistical significance be made about the mean subjective responses, it is necessary that a probability model about a clearly defined population be established before the test; this is discussed in Sections 2 and 4 below. It is also necessary that the codecs be assigned at random to the many scene-test-location combinations; this is considered in Document Number T1Q1.5/89-219 and discussed in Section 3 below.

It is desirable to design the test procedure of such a structure and size that the results enable a clear winner to be declared with prescribed probability. Section 4 describes how this can be done. However, the structure and size may not be acceptable to the Committee, and an alternative design and analysis is described in Section 5 that yields statistical significance or confidence statements to be made but does not guarantee a clear winner. Recommendations are made in Section 6.

2. Population of Observers

Section 6 of Document Number T1Y1.1/89-060R2 permits very broad selection of viewers, "both expert and naive viewers" but not "experts on digital image compression." Perhaps other examples of permissible types should be given: men and women, age limits (?), etc. There should be no question whether the different testing laboratories have different ideas about what population they are selecting from. All reviewers should be thoroughly trained in "warm-up" tests so that there will be no missing data in the formal tests.

3. Randomization

The randomization techniques in Document Number T1Q1.5/89-219 are presumed to be satisfactory, but it is not clear whether different random orders will be used at the different testing laboratories. They should be. For example, if the same codec was presented first in the basic quality test at all testing laboratories, perhaps it would tend to get a low score simply because it was first. This might tend to be averaged out if it had other positions on other tests, but it is better not to incur the possibility, especially if the basic quality test is given more weight than others. (Since all of the tests seem to have a substantial element of random variation in them, not departing much from unity in the weights is preferable, since unit weights give the maximum ability to average out random variation if all tests have the same standard deviation of random variation. However, aside from this, the choice of weights lies in the area of the subject specialist rather than that of the statistician.)

4. Design of a Selection Procedure

Under certain assumptions it is possible to design a test (experiment) so that one has a prescribed probability of selecting a clear winner (Encyclopedia of Statistical Sciences, S. Kotz, N.L. Johnson, and C.B. Read (eds.), Vol. 8 (1988), pp. 337-345, "Selection Procedures" by J.D. Gibbons, John Wiley & Sons, abbreviated ESS hereafter).

One first has to be clear as to what population of observations one is talking about. For this reason it appears that the same group of observers should be used in all sessions at each testing laboratory, contrary to the last sentence on page 10 of Document Number T1Y1.1/89-060R2. There may be difficulty in getting everyone present at the first session back to all subsequent sessions, but the importance of it should be emphasized because in the probability (statistical) model seen as necessary, the individual observation will be the average (or sum), possibly weighted, of the scores of the individual observer on all tests. Hence each observer must be identified, say by name, so that his scores can be combined. Some initial observers will be lost by sickness, unavoidable conflicts, or lack of interest, but a sufficiently large initial group, say 30, should be obtained such that the losses will not be crucial. It is further proposed that each evaluation laboratory randomly reduce (by blind randomization procedure) the number of observers from the testing laboratories so that each testing laboratory has the same number. This will preclude undue influence by any one testing laboratory.

The reason for using the same group of observers throughout may become clear from a summary of the selection procedure. The procedure is based on the choice of an "indifference zone." It is assumed that one would be indifferent as to the codec selected from k candidate codecs if two or more true mean scores μ_1, μ_2, ..., μ_k differed by as little as a prescribed value, say δ, which has to be picked by the committee. For example, on the double-stimulus continuous-quality scale of Appendix 6 (Doc. 89-060R2) ranging from 0 to 100, the committee might choose δ=3. The committee must also choose the probability, P, of selecting the

best codec if indeed there is one that has a true mean score, μ, greater than all the others by at least δ=3. For example, the committee might choose P=0.95.

There is one additional number needed in order to determine how many observers, n, are needed in a one-stage test to be able to select a unique winner (better by δ=3) with probability P=0.95. Unfortunately, it is not a number the committee is at liberty to specify, it must <u>know</u> (or assume it knows) the (true) standard deviation σ of the population of individual observer mean scores. Then the required total sample size (total over all testing laboratories) is given by

$$n = (\sigma\tau/\delta)^2$$

where τ depends on the number of codecs, k, and probability, P, and is given in Table 1, page 339 of ESS. If k = 4 and P = 0.95, then τ = 2.92. (There is an assumption of normally distributed observations, but that should be easily satisfied by the individual mean scores.) For example, if σ = 5, then in the example above, $n = (5 \times 2.92/3)^2 = 23.7$. Thus, if 3 groups of 8 observers each produced individual mean scores for the 4 codecs of

$$x_{111},\ldots,x_{118} \; ; \; x_{121},\ldots, x_{128} \; ; \; x_{131},\ldots, x_{138}$$

(x_{138} is response for codec 1, group 3, observer 8)

$$\vdots$$

$$x_{411},\ldots, x_{418} \; ; \; x_{421},\ldots, x_{428} \; ; \; x_{431},\ldots, x_{438}$$

and the corresponding 4 codec sample means $\bar{x}_1..$, $\bar{x}_2..$, $\bar{x}_3..$, $\bar{x}_4..$, ($\bar{x}_1..$ being the sum of the 24 numbers in the first line above divided by 24, for example) were calculated, the largest would identify the clear winner (by δ=3 with probability P = 0.95).

In practice it is difficult to know σ unless a substantial amount of experimentation has already been done in exactly this format. Theory provides for this by asking that two stages of testing be allowed for; that is, the entire test might have to be repeated except that the number of observers may have to be different. The trick is that σ is estimated by the sample standard deviation, s, [ESS, Vol. 5, page 681, eq. (7)] from n_0 initial observations on each codec. Then we would calculate the number $n_1 = 2\,(sh/\delta)^2$ where h depends on k and n_0 and is given in Table 2, page 340 of ESS. If $n_1 \leq n_0$, no further testing is needed. If $n_1 > n_0$, then a second stage of testing with $n_1 - n_0$ observers is needed. For example, if n_0 = 24 and s turns out to be 5, then h = 2.09 (for k=4) and

$$n_1 = 2\,(5 \times 2.09/3)^2 = 24.3.$$

Hence, rigorously, one would have to conduct a second stage of testing with one observation on each codec and calculate the $\bar{x}_i..$ from all 25 observations on each codec to declare a winner with the prescribed assurance. In practice, if the first stage came this close, one might well let his prescription slip a bit and be satisfied with the one stage.

5. Analysis by Multiple Comparisons

Realistically, the committee does not know what σ is, and it may feel that a two-stage experiment is impractical. An alternative is to accept that a clear winner may not emerge and to be satisfied with simultaneous confidence intervals on the $k(k-1)/2$ mean differences $\mu_i - \mu_j$.

Before going into this in detail, the concept of confidence interval will be introduced in terms of a single mean value. Suppose a sample of n independent observations are made on a normal distribution with (true) mean μ and (true) standard deviation σ. The sample mean, $\bar{x}$, and sample standard deviation, s, are calculated. Then $\bar{x}$ is an estimate of μ but with some uncertainty. It is possible to calculate from $\bar{x}$ and s a confidence interval that indicates the extent of that uncertainty if P, the probability that the interval will include μ, is specified by the experimenter (or committee). If $P = 0.95$, he will be "95% confident" that the interval includes μ, which means that if he does this routinely on test after test, about 95 out of 100 intervals will cover the corresponding μ and 5 will not but no one will know which 5. The confidence intervals in this case are calculated as $\bar{x} \pm tsn^{-1/2}$, where t depends on P and n (t can be found in a table of "Student's t").

Here there are several codecs, say k as above, and they are compared by estimating the differences in true means, $\mu_i - \mu_j$. With $k = 4$ codecs there are $4x3/2 = 6$ differences, $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, ..., $\mu_3 - \mu_4$, which can be estimated by the sample means $\bar{x}_1.. - \bar{x}_2..$, $\bar{x}_1.. - \bar{x}_3..$, ..., $\bar{x}_3.. - \bar{x}_4...$ Confidence intervals for the true differences can be calculated, but the committee would have to specify a probability P that all six intervals include the respective differences $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, ..., $\mu_3 - \mu_4$ (because it is not known beforehand which $\bar{x}_i..$ will be the largest). Then the confidence intervals are given by

$$\bar{x}_i.. - \bar{x}_j.. \pm qsn^{-1/2}$$

where s is the common sample standard deviation as in Section 4 above and q is a tabulated value depending on P, k, and n [ESS, Vol. 5, page 682].

For example, suppose $k = 4$ and the codec sample means turned out to be $\bar{x}_1.. = 68$, $\bar{x}_2.. = 70$, $\bar{x}_3.. = 75$, $\bar{x}_4.. = 63$, based on 3 groups of 8 observers, so $n = 24$. Then codec No. 3 is the apparent winner, but is it significantly better than the others? Suppose also, as in Section 4, that the common sample standard deviation s turned out to be 5 and that the committee had prescribed a P of 0.95. Then q is about 3.78 so $qsn^{-1/2} = 3.78 \times 5 \times 24^{-1/2} = 3.86$. The confidence interval for $\mu_3 - \mu_2$ is

$$5 \pm 3.86 = (1.14, 8.86),$$

so codec No. 3 is significantly better than codec No. 2, and the other two are even farther behind. In general, what is required for a clear winner is that the k - 1 confidence intervals involving the largest sample mean not extend into negative values. If s had been 7 rather than 5 above, then $qsn^{-1/2} = 5.40$, and No. 3 would not have been a clear winner.

6. Recommendations

These recommendations being proposed should be taken as a basis for discussion.

(1) Two stages of testing seem impractical, so that the analysis by multiple comparisons of Section 5 should be specified for the analysis laboratories. The Committee should pick probability P as described in Section 5, before testing begins.

(2) Each testing laboratory should assemble a specified substantial number, say at least 30, of observers who can be committed for all viewing sessions and should be identified. Besides being necessary for the probability model (without which a statement of the statistical significance of the results cannot be made), this has the advantage of eliminating the need for further training before each subsequent session.

(3) All record sheets should be sent to one analysis laboratory and copies to the other. Each laboratory will eliminate observers for which the record is not complete over all scenes, tests, and codecs. Any excess number of observers of any testing laboratory over the minimum number will be discarded by random selection, so that all contribute the same number. Each laboratory will calculate a mean score for each of the remaining n observers for each codec. From these a mean for each codec and an estimate of the (assumed) common standard deviation based on k(n-1) degrees of freedom needed in recommendation (1) above can be calculated. Then the confidence intervals involving the apparent winner given in Section 5 can be calculated to determine whether it is significantly better than the others. The foregoing calculations could be modified to include the possibility of interlaboratory differences.