THE SELECTION OF RETURNS FOR AUDIT BY THE IRS

John P. Hiniker, Internal Revenue Service

BACKGROUND

The Internal Revenue Service, hereafter referred to as the IRS, is responsible for administering the Internal Revenue Code as passed by Congress. In fulfilling its responsibilities, the IRS has many on-going programs, including Returns Processing, Taxpayer Service, Examination, Appeals, Criminal Investigation, Collection, etc.

One of the most important of these programs, in terms of resources expended and taxpayer impact, is the examination program. The main purpose of the examination (or audit) program is to help ensure a high degree of voluntary compliance with the Federal tax laws. In attempting to accomplish its purpose, tax returns of all types are examined and corrected, if necessary. The percentage of returns examined has not varied substantially over the past number of years (approximately 1-2% of filings) and, due to limited audit resources, will always be only a relatively small fraction of those filed. This limitation of resources makes it imperative that the returns selected for examination be those with the greatest impact in deterring non-compliance.

The general public is most familiar with the individual tax returns--Form 1040EZ, Form 1040A or Form 1040--which are filed yearly. During FY 1986, approximately 102.2 million individual tax returns were filed. Also during FY 1986, approximately 1.1 million individual returns were audited (generally corresponding to 1984 and 1985 filings). In terms of individual filings, the IRS strategy for selecting returns with the greatest impact on encouraging voluntary compliance is to select returns with a high probability of significant tax change. This is done within categories of return filings and geographically to ensure broad coverage.

Not considered audits by the IRS are communications with taxpayers regarding relatively simple and readily identifiable problems that can be resolved easily. Most of these relate to items on the return, identified manually and by computer, which appear to be unallowable by law. For example, claiming gasoline tax paid as a deduction on Schedule A, which is not allowed under present law.

On the other hand, there are many issues that are not readily identified or easily resolved. These do require the thoroughness of an actual examination. The reasons associated with being selected for audit are many and include third party information being received (e.g., Form 1099's and W-2 statements), claims for refund, special enforcement efforts, related returns being examined (i.e., partnerships, prior-year returns, etc.) and, to a limited extent, pure random selection such as that done for the Taxpayer Compliance Measurement Program. The largest number of returns (approximately two-thirds), however, are selected under the regular examination program.

In 1960, before the IRS was able to utilize automatic data processing in its regular examination program, individual returns were selected by manual review, which attempted to identify the returns most in need of audit (i.e., returns with high probability of significant tax change). Criteria based on experience were utilized. As can be imagined, the task was monumental, with not all returns reviewed and with a lack of uniformity on those that were.

With the advent of automatic data processing, these subjective criteria were formalized and programmed, with all individual returns being screened. A vastly greater number of returns were identified than could be audited, which again required manual review with problems of uniformity and coverage. The system was further refined to rank the selections by number of separate criteria met (i.e., multiple criteria method). This was a much better approach, but it assumed that all criteria were objective, independent, and of equal value, which was not the case.

The advent of automatic data processing did, however, allow for other, more sophisticated, approaches to be considered in the selection process. The one this paper considers is the Discriminant Function Approach.

DISCRIMINANT FUNCTION APPROACH

The discriminant function approach was first developed by R.A. Fisher in the 1930's.[1] It essentially reduces a multivariate situation (in this case the variables reported on a Form 1040 or Form 1040A) to a single variate (score), which can be used to classify an observation into one of two (or more) populations; i.e., "need-to-audit" population or "no-need-to-audit" population. Its usual form would look like:

$$Z = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 .. \lambda_i X_i .. \lambda_n X_n = \sum_{i=1}^{n} \lambda_i X_i$$

where i= 1,2,3...n represents the variable on the return being utilized in the formula,

λ_i= a coefficient developed through analysis of the statistical distribution for the i^{th} variable,

X_i= the value reported on the return for the i^{th} variable, and

Z= the resultant score assigned to the return.

Thus, once the λ_i values are determined through mathematical analysis and given the X values reported on a tax return, a Z score can be computed for the return which would allow for that return to be classified into either the "need-to-audit" group or the "no-need-to-audit" group.

The discriminant function approach is a classification technique, not a ranking technique, although IRS's experience and testing has indicated it can be used effectively for both purposes. A high degree of correlation has been found to exist between the resultant score and subsequent tax change, enough for the IRS to feel that the score suffices as a ranking device by which returns can be ranked as to the probability of high tax change. Thus, the IRS is in a position to optimize the selection of returns by taking only the highest scored returns commensurate with any given level of audit resources.

Most of the work done on this approach in the past has been done for two (or more) populations in which measurements for the variables were assumed to be normally distributed with equal covariance matrices.

The covariance assumption of normality of the variables would imply that the resultant combinatorial value (or Z value) would also be normally distributed, which has its advantages. In the IRS situation, all the possible distributions of variables are not normally distributed. There are discrete variables and continuous variables which have truncations and skewness. However, experience indicates the technique will "take" significant departures for normality (i.e., it is robust).

Under the second assumption, that of equal covariance matrices, the optimum discriminant function turns out to be linear. However, if the restriction of equal variances and covariances is removed, the optimum function is quadratic. If this is the case, the "best" linear function can be developed, even though it is not the optimum. This can be used or the quadratic function can be used. In terms of this discussion, equal covariance matrices implies:

- the variance for variable i in the need-to-audit group is the same as for i in the no-need-to-audit group; and
- the relationship (or covariance) between variable i and variable j in the need-to-audit group is the same as between variable i and variable j in the no-need-to-audit group.

The assumption of equality of covariance matrices allows for the pooling of these matrices in the derivation of the lambda (λ) values.

The actual derivation of the λ values is straightforward and described in many statistical texts. Tests of the efficiency of the classification function are also readily available.

Although well defined in literature, the discriminant function approach required a number of modifications to fit IRS' situation. These modifications moved the approach from a highly theoretical classification model into a practical application with enormous impact on the way IRS performs its functions.

Taxpayer Compliance Measurement Program (TCMP)

A suitable data base for research and formula development was available through the Taxpayer Compliance Measurement Program. This program represents in-depth audit results for a probability sample of all individual taxpayers filing returns. The data base contains taxpayer reporting data, along with the results of auditing the return.

TCMP is conducted every three years in the individual tax return area and provides the IRS with the means to monitor compliance levels,as well as providing a data base for various compliance studies. It thereby allow for periodic updating of formulas, as well as the initial development.

Since the IRS generally allocates its audit resources on the basis of different classes of taxpayers (audit classes), the TCMP data are structured by these audit classes. This is reasonable in that these classes represent specific types of taxpayers with their own compliance patterns and different tax return line item data available. (These classes are generally defined on the basis of income and presence or absence of farm or nonfarm business income.)

For purposes of formula development, each audit class was considered independently, with a formula developed for each class. The fact that the resulting formulas were different

attests to the desirability of developing specific formulas for each audit class. Thus, there is no single formula being utilized by the IRS in the selection of returns for its regular audit program.

Determination of the Two Populations of Interest

Although, generally, the IRS has defined the two populations of interest as the need-to-audit group and the no-need-to-audit group, an exact definition was based on an analysis of the various populations involved. This analysis indicated that not two -- but many -- subpopulations existed, each having different characteristics. This would ordinarily suggest that some other multivariate technique might be more appropriate. The IRS, however, chose to continue using the basic discriminant function approach, largely due to the possibility of using the score as a ranking scheme, which was basically the goal.

An analysis of the various populations involved was carried forward by dividing the research file of completed audits (for an audit class) into eight mutually exclusive and exhaustive groupings based on the audit results. These preliminary groupings were as follows:

- tax decrease of $100 and over;
- tax decrease under $100;
- no tax or variable change;
- no tax, but variable change;
- tax increase less than $50;
- tax increase $50-$99;
- tax increase $100-$199; and
- tax increase $200 and over.

Traditionally, the IRS has defined the need-to-audit group as consisting of those taxpayers whose returns, if audited, would result in a significant tax change. The IRS attempts to be even-handed in this, giving just as much importance to a tax decrease (or refund) as to a tax increase. However, the possibility existed that these tax decrease returns may have different characteristics than the tax increase returns and to combine them would lessen the effectiveness of any formula developed by neutralizing reporting differences that each group might have. This point was, and still is, not generally understood. There is a tendency to view the process of defining the two groups as a simple reflection of the objectives, without consideration of various sub-populations that may exist in the data.

Profiles (generally, the average and variance of the amount reported for each variable) were developed for each of the eight grouping from the research file. These were compared with the general conclusion that there existed basically four different subsets for every audit class. These four are as follows, with the X and Y varying by audit class:

Subset 1. -- tax decrease $X or more;
Subset 2. -- no tax change or tax increase or decrease less than $X;
Subset 3. -- tax increase $X to $Y; and
Subset 4. -- tax increase $Y and over.

At this point in development a seeming dilemma existed, with four distinct subsets emerging and an approach relevant to only two. However, the third subset had a general variable reporting profile that was between Subset 2 and Subset 4, which suggested that no matter what was done by way of assigning them to one group or another for development purposes, their final scores -- as well as the final scores for all tax returns filed in the future for this "subset" -- would fall between those of Subset 2 and Subset 4. Consequently, they were dropped from further consideration in the development of a formula. Later testing and actual usage of the formulas bore out the initial determination that these returns do, in fact, tend to distribute themselves between the no change subset and the high change subset.

The first subset, that of tax decreases, posed a more serious problem. Analysis of their profile indicated that they had no unique set of filing characteristics (as evidenced by large variances and variable averages appearing to be randomly distributed in relation to the other groups). This suggests that they could not be effectively distinguished as a group, even with a separate formula, and would tend to distribute themselves independent of any definition used. They were subsequently dropped from further consideration in the development of the formula.

Later testing of the formulas developed using only Subsets 2 and 4, but applied to all subsets, bore out the fact that these tax decrease returns distributed themselves all over the range of possible scores and it was not prejudicial to "refund-type" taxpayers to omit them from the formula development. However, over time, this determination became politically unacceptable to the administrators, so that current definitions of groupings are made without consideration of whether the audit resulted in a tax increase or tax decrease (i.e., on an absolute tax change basis). Some loss

of efficiency can be expected with this kind of decision.

Determination of the two basic populations of interest was a major effort. A major point to be emphasized is that the data dictate the definitions of the populations of interest as much as the administrative purposes of the formulas.

These two populations -- no tax change or tax change less than X amount (π_2) and tax change increase greater than Y amount (π_1) -- were used in subsequent efforts. The X and Y amounts varied by audit class breakout.

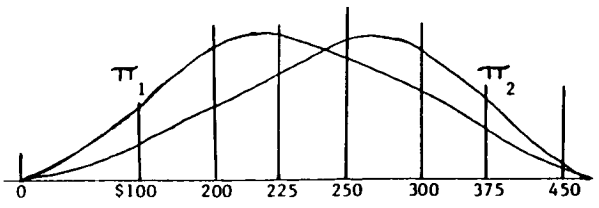
The Likelihood Ratio Transformation

It is readily apparent, in an application such as IRS', that nice, neat, normal distributions of the variables do not exist. As can be imagined, distributions utilizing income data were extremely skewed to the right. Many distributions are truncated or discrete. Although it is generally true that the discriminant function approach is robust, significant departures from normality exist. Still, rather than consider possible non-parametric type approaches to the classification/ranking problem, IRS instead chose to consider possible transformations to the data to lessen the impact of non-normality. A great deal of effort was expended exploring various transformations (log, square root, etc.) to at least bring the distributions into more compact, single moded form. These efforts were only moderately successful.

In order to bring uniformity and simplicity into this hodge-podge of different types of variables and distributions, a single transformation of variable response was developed, using the concept of likelihood ratio. That is, suppose for any given variable, the density distribution of response, for the first population, π_1, was overlapped with that of the second population, π_2. (See Figure 1.) Then suppose the

Figure 1.-- Hypothetical Example of Likelihood Ratios for Variable X

Population π_1--High tax increase (need-to-audit)
Population π_2--Little or no tax change (no need-to-audit)



Note: As can be easily visualized, the less the response (taxpayer reported amount), the more likely the return belongs to Population π_1. The likelihood ratio reflects this.

scale was partitioned into zones of response using as many zones as necessary in attempting to depict the true distributions from sample data. In each of these zones, there would be a ratio of response corresponding to the likelihood of the individual being in a particular population.[2] This can be demonstrated by the following hypothetical example (in Figure 2).

In the case of discrete data, such as filing region, the likelihood ratio transformation would be determined basically the same way, --see Figure 3.

As exemplified there, a return filed in Region D is much more likely to belong in π_1, than a return filed in Region F. The transformed response reflects this (2.25 vs. 0.47).

Thus, "zones" were defined for each variable and a likelihood ratio for each zone was determined from the development data. In theory, if the variable distributions are completely defined, there could be a zone for every possible response, which is generally the case for discrete variables. However, with continuous variables being considered, and using sample data, the distributions

Figure 2.--Hypothetical Values for a Continuous Variable

ITEM	HYPOTHETICAL VALUES OF VARIABLE (X)							
	0-$100	$101-$200	$201-$225	$226-$250	$251-$300	$301-$375	$375-$450	$451 AND OVER
π_1 FREQUENCY	10%	25%	20%	10%	20%	10%	4%	1%
π_2 FREQUENCY	6%	15%	8%	10%	25%	21%	10%	5%
$\frac{\pi_1}{\pi_2}$ LIKELIHOOD RATIO	$\frac{10\%}{6\%}$	$\frac{25\%}{15\%}$	$\frac{20\%}{8\%}$	$\frac{10\%}{10\%}$	$\frac{20\%}{25\%}$	$\frac{10\%}{21\%}$	$\frac{4\%}{10\%}$	$\frac{1\%}{5\%}$
TRANSFORMED RESPONSE	1.67	1.67	2.50	1.00	0.80	0.48	0.40	0.20

Figure 3.--Hypothetical Values for a Discrete Variable

Item	Region						
	A	B	C	D	E	F	G
FREQUENCY π_1	10%	13%	15%	18%	20%	8%	16%
FREQUENCY π_2	16%	10%	18%	8%	15%	17%	16%
$\frac{\pi_1}{\pi_2}$ LIKELIHOOD RATIO	$\frac{10}{16}$	$\frac{13}{10}$	$\frac{15}{18}$	$\frac{18}{8}$	$\frac{20}{15}$	$\frac{8}{17}$	$\frac{16}{16}$
TRANSFORMED RESPONSE	0.63	1.30	0.83	2.25	1.33	0.47	1.00

tended to be so "jagged" due to sampling error that around ten zones seemed to describe the distributions as well as could be expected.

The likelihood ratio for the zone is equivalent to expressing the odds for a return being in π_1 or π_2 (assuming equal sizes). Thus, if the transformed response for a variable response is less than 1.00, the variable response is associated with π_2. If the transformed response is greater than 1.00, the variable response is associated with π_1.

A tape file was created with the appropriate transformed response replacing the original reported variable response on each return.

A further reduction of skewness existing in the distributions of transformed responses could be accomplished by using the log of the likelihood ratio or its probability equivalent. However, this was not done. (While both the log expression and the probability expression were felt to be better for classification, the feeling did not carry over into the ranking objective of the effort.)

To this point in development, data other than summary type was unnecessary. It is interesting to note that all one would have to do is take new returns and convert the reported amounts to the corresponding transformed responses and add to get a reasonably effective classification system. The higher the total sum of transformed responses, the more a return would be associated with π_1. However, there would be no adjustment for correlation between, the variables. Since, at least in IRS' case, these correlations are known to exist, they had to be adjusted for. In applications where the variables are known to be independent, it may be unnecessary to perform the standard mathematics of the approach.

Using the standard discriminant function mathematics [3], and the transformed data tapes, the appropriate equations were developed and solved to yield λ_i values. At this point, however, many variables in the formula were not contributing to the effectiveness of the formula. The contribution was of each variable was determined and the variable making the least contribution was dropped. This was done successively until deletion of another variable tended to make a difference in formula effectiveness (generally, this occurred at between 10-15 variables).

The Look-up Table Concept

As stated earlier, the usual form the approach takes is:

$$Z = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \ldots + \lambda_i X_i + \ldots + \lambda_n X_n$$

which suggests that reported variable responses are multiplied by the appropriate λ_i's and added to form the score. However, by using zones of response and likelihood ratios, it would appear that we have complicated the problem by first having to relate each variable response to its appropriate likelihood ratio transformation and, then, multiply the transformed value by the λ value for the variable and, then, sum for all variables.

Since the likelihood ratio is known for each zone, and the λ value is similarly known for each variable, it is possible to multiply the two in advance, thereby eliminating the need for continuously multiplying. In other words, it is possible to construct a look-up table for each variable, which would reflect the likelihood ratio of the zone times the λ value. While current advances in computerization make this a trivial reduction in arithmetic, it wasn't at the time. Further, it offered a simple approach to handling zeroes and extreme values. Thus, in (operational) usage, the computer would take the reported amount for variable i, refer to the look-up table for variable i and take as the contribution to the total score the points corresponding to the reported amount claimed. In reference to our earlier hypothetical example in computing the likelihood ratios, the table for a continuous variable might look like the following

if the derived λ_i value was 2.0.

Figure 4 -- Hypothetical Look-up Table for Continuous Variable X with $\lambda_i = 2.0$

Reported Amount	Contribution
$0 - $100	3.34
101 - 200	3.34
201 - 225	5.00
226 - 250	2.00
251 - 300	1.60
301 - 375	.96
376 - 450	.80
451 & up	.40

Thus, in effect a formula was developed that would allow for future return filings to be scored.

Second Stage Function

It was mentioned earlier that the discriminant function approach is a classification approach and not a ranking approach. However, the IRS was not interested in setting a cut-off score with its associated Type I and Type II errors. IRS concern was really only in the upper rankings of scores, where returns would be selected for audit. In fact, IRS only wants to select returns with the very highest potential of those tax returns that might be classified as need-to-audit.

Since these very high tax increase returns represent only a small percentage of cases, IRS was concerned that the development of the formulas might be dominated by cases only slightly above the cut-off point for high tax increase cases. One approach to this ranking problem of the higher potential tax change cases would be to use the discriminant function to separate the two populations and, then, to use multiple regression on the expected high tax increase population, to estimate the amount of tax change. This would result in a ranking of returns by expected tax change. IRS did not go this way, although it is now in the process of exploring this approach as an alternative.

Methodology. -- The approach taken, largely due to having the discriminant function programs in place, was to develop a second stage discriminant function formula to be used in conjunction with the first. This was done by using the first formula to classify returns as to being in the high expected tax increase population. Only those returns classified as high expected tax increase were utilized in developing a second formula along the same lines as the first.

The cut-off point was determined on the basis of having enough π_1 and π_2 type returns to allow for development of another formula. If the initial formula was reasonably efficient, relatively few π_2 type returns would be expected in the upper scores. This was generally true and would not have left enough π_2 type returns except that π_2 usually contained three times the number of cases as π_1 (which reflects the generally high level of compliance found in auditing returns). A cut-off point between the 40th and 60th percentiles was generally found to be acceptable in terms of both development sample sizes and consideration of the associated Type I and II classification errors. Thus, all returns in the development file with determined discriminant function scores below a certain point were discarded and another discriminant function formula was developed using only these π_1 and π_2 returns with scores above the cut-off point.

In effect, it was concluded that a large grouping of filed returns is easily classifiable and holds little potential for tax change. However, because of their great number, these returns influence the development of the initial formula and IRS wanted to remove that influence. It was felt that a two-stage formula would do that, and give better discrimination in the range of returns where workload was drawn. Using this two-stage approach, the correctness of ranking among the highest scores was improved approximately 10-15%.

Theoretically, a third, fourth, etc., stage approach would allow for further improvement, but after the second stage there were not enough remaining little or no-tax change cases to allow for formula development.

Implementation. -- Ordinarily, a two-stage approach would also require a two-stage implementation approach in ranking new returns filed. That is, the first stage would classify and the second stage would rank. It turned out that the rankings in the upper end were not changed by utilizing the second-stage formula by itself. This means using just the second-stage formula (as a one-stage system) to rank new filings was equivalent to using both stages in the upper range of scores. As such, IRS was able to avoid the additional complexity of a two-stage approach in its operational use of the formulas.

Initial testing of the discriminant function approach was made by scoring all returns in the data file, including those discarded due to being tax decrease or medium tax increases. The formula scores verified the assumptions made during early development.

Secondary testing was made by selecting 13,000 unreviewed returns from a previous filing year and determining which returns would be audited if 10% were to be audited using the formula for

selection. The criteria method and the multiple criteria method (as described in the beginning of this paper), as well as a manual review method, were applied to the same set of returns and given the same instruction. All returns selected by any method were audited and the results compared. The discriminant function approach was superior in maximizing the identification of high tax change returns.

Full implementation of this discrimination function approach took place in 1971 and, with periodic formula updating, has been utilized ever since.

CLOSING REMARKS

The experience described in this paper occurred approximately 20 years ago. The approach has been enhanced to some extent, but further research is still likely to be profitable. The IRS has tried to improve upon the system by contracting out for development using other approaches but, for the most part, alternatives have turned out to be not as effective.

Thus, the basic approach, as discussed in the paper, serves as the principle means of selecting individual tax returns for audit by the IRS. As such, it clearly demonstrates how statistical techniques can be utilized to solve very practical operational problems.

NOTES AND REFERENCES

The opinions and conclusions expressed in this article represent those of the author and do not necessarily represent the position of the Internal Revenue Service.

[1] Fisher, R.A. "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol. 7 (1936), pp. 179-188.

[2] Actually, the ratio of response would correspond to the likelihood ratio (or odds) of an individual being in a particular population only if the populations were of similar size, which was not necessarily the case.

[3] Kossack, Carl F. "On the Mechanics of Classifications," Annals of Mathematical Statistics, vol. 16 (1945), pp. 95-98.