Tilburg University

Audit assurance model and Bayesian discovery sampling

van Batenburg, P.C.; Kriens, J.; Lammerts van Bueren, W.M.; Veenstra, R.H.

Link to publication in Tilburg University Research Portal

ERS/TY

UNIVERSITEIT
BRABANT

POSTBOX 90153
5000 LE TILBURG
THE NETHERLANDS

DEPARTMENT OF ECONOMICS

RESEARCH MEMORANDUM

AUDIT ASSURANCE MODEL AND BAYESIAN DISCOVERY SAMPLING

P.C. van Batenburg, J. Kriens,
W.M. Lammerts v. Bueren, R.H. Veenstra

FEW 471

Audit Assurance Model and Bayesian Discovery Sampling

Objections to the Audit Assurance Model from audit theory and
statistical methodology

and

Bayesian Discovery Sampling, a better method to utilize the
auditor's 'professional judgement' in sampling

P.C.van Batenburg[1], J.Kriens[2], W.M.Lammerts van Bueren[3]
and R.H.Veenstra[4]

1)   Senior Statistician at Touche Ross Nederland Center for Quantita-
     tive methods and Statistics.
2)   Tilburg University, The Netherlands, and advisor to the Touche
     Ross Nederland Center for Quantitative methods and Statistics.
3)   Rotterdam Erasmus University, The Netherlands, and advisor to the
     Touche Ross Nederland Center for Quantitative methods and
     Statistics.
4)   Partner and National Director of Auditing of Touche Ross
     Nederland.

Author's address:

TRN Groep / CKS
Westhage Offices,
Anthony Fokkerweg 1,
NL 1059 CM Amsterdam,
The Netherlands.

I. Introduction to both parts

I.1. Stating the problem

Auditors often use statistical sampling to confirm their preliminary
assessments of the quality of a population, expressed as the error
fraction in the financial statements to be audited. These assessments
are based on experiences in the past (previous audits) and on audit
activities in the present, such as the review of the system of internal
controls, analytical review and compliance tests.

A statistical sample, however, will not result in the determination of
the exact population error fraction. Instead, an interval is specified
that will include the unknown real error fraction up to a certain ex-
tent: the <u>confidence</u> level. The width of the interval determines its
<u>inaccuracy</u>. Given the sample size, the inaccuracy can only be improved
at the cost of the confidence level, and vice versa.

In the last few decades, the sizes of the populations to be audited have
grown, resulting in a necessity to reduce sample inaccuracy of the error
fraction (to keep the inaccuracy in monetary units small enough). On the
other hand, the pressure on audit costs has made a reduction of sample
sizes unavoidable. Therefore, auditors and statisticians have been
(re-)searching for methods that combine confidence levels the statis-
tician can agree with, inaccuracy levels the auditor can depend on, and
sample sizes the client can pay.

These methods can be classified into two categories:

> The Audit Assurance Model claims to asses the <u>confidence</u> level
> required for a particular audit sample in order to reach a requi-
> red level of overall assurance, and based on specified levels of
> inherent assurance, assurance from analytical review and assurance
> derived from compliance tests.
> The auditor substitutes the statistical 'confidence level' by (the
> not very clearly defined) 'overall assurance'. The general idea
> behind overall assurance amounts to the certainty that the auditor
> will find any material error, using a mixture of his knowledge and
> skills and the sample results.
>
> Bayesian statistical methods enable to influence the <u>inaccuracy</u>
> level of a statistical test by using existing information.
> Prior information and knowledge, resulting from the auditor's
> general experience and specific work, is quantified in the form of
> a probability distribution of possible error fractions in the
> population to be sampled. Assuming the 'correctness' of this in-
> formation, the sample size that is necessary to reach the inaccu-
> racy level that was originally requested is smaller than the
> 'classically' determined sample size.

In part two, the authors will discuss their objections against the Audit Assurance Model, both from an auditor's and from a statistician's point of view. It is the authors' goal to show that the Audit Assurance Model is:

- formulated in quantities that affect the auditor's confidence level, but should affect his inaccuracy level;
- using statistical assumptions that can not be verified;
- giving unacceptable (though methodologically consistent) results once these assumptions have been dropped.

In part three, the authors present a Bayesian alternative to overcome these drawbacks. In this method, the auditor uses last year's audit sample results to specify a probability distribution of the error fraction in last year's audit population. Then, the auditor uses his knowledge about inherent quality, accuracy from analytical review and accuracy derived from compliance tests, to specify to what extent this probability distribution can be considered as prior information about this year's error fraction. This depends on his assessment of 'the stability of accounting processes'.

The sample size that finally remains in order to reach the acceptable inaccuracy level, determined by the auditor's materiality conditions for this year's audit population, will often be much smaller than classical sampling theory would yield.

First, the 'classical' way in which auditors determine sample sizes, is described.

I.2. Discovery sampling

Discovery sampling is a method to derive the size of an audit sample (n) from population size (N), intolerance fraction (p_1, the auditor's materiality divided by population size) and the maximally tolerated probability (β_o, sampling risk, the complement of confidence level) that a population with ('intolerable') error fraction p_1 or more yields a sample that suggests a lower ('tolerable') error fraction. Roberts (1978) defines discovery sampling as: 'a procedure for determining sample size required to have a stipulated probability ($= 1-\beta_o$, aut.) of observing at least one occurrence ($=$ error, aut.) when the population occurrence rate is at a designated level ($= p_1$, aut.)'.

Statistically it is based on the fact that the number of errors in such a sample, $\underline{k}$, (random variables will be underlined in this paper) follows a hypergeometric distribution. For relatively large populations (such as when population and materiality are expressed in monetary units), this distribution can be approximated by a binomial distribution. From the resulting number of errors, the upper limit of a confidence interval for the unknown population error fraction is calculated. When no errors occur, this upper limit should equal the materiality fraction. The method by which such an upper limit can be found is explained, for example, by Blyth (1986). The upper limit is the smallest value p* for which:

$P(\underline{k} = 0 \mid N, n \text{ and } p*) \leq \beta_o$.

This upper limit should equal the auditor's materiality fraction p_1 when $k=0$ in a sample of size $n*$, for which:

$P(\underline{k} = 0 \mid N, n* \text{ and } p_1) \leq \beta_o$.

Using binomial probabilities, the minimal sample size $n*$ can be found from:

$(1-p_1)^{n*} = \beta_o$ so $n* = \log(\beta_o)/\log(1-p_1)$.

> (To attain an integer value for $n*$, the numerical result is always rounded up.)

Further approximation of the binomial probability function by a Poisson distribution (often used in Anglo-Saxon oriented audit firms) yields a very easy rule of thumb for calculating the necessary sample size:

np_1=constant (being for example 3 if $\beta_o = 0.05$).

II. Objections to the Audit Assurance Model from audit theory and
     statistical methodology

Part two is organized as follows. Section 1 formulates the Audit Assu-
rance Model (AAM from now on) and its influence on the way auditors use
discovery sampling. In that section, the AAM is criticized from audit
theory and we show our statistical objections against the AAM. In sec-
tion 2 it is shown, mathematically and by means of graphs, that the
'statistically improved' AAM will give unattractive outcomes to the
auditor. Section 3 concludes.

II.1. The Audit Assurance Model

II.1.1. Description of the model

The AAM has appeared in many different forms. Bailey (1981) presents 4
slightly different models, with the same objective (quoted from Bailey,
page 231): 'the linkage between various compliance and substantive
tests of details together to render a combined reliability measure'.
Each of them can be reformulated into:

$$OA = 1 - \beta_o(1-A),$$

in which:

OA= the level of <u>overall assurance</u> to be attained. Overall assurance
    is the certainty that the auditor will not miss a material error
    (an error which magnitude is at least the intolerance fraction p_1)
    in his audit;
A= the level of <u>assurance</u>, which means the certainty the auditor has
    that material errors will either not be present or will have been
    detected before the population is subjected to sampling;
β_o= the sampling risk: the probability that a population with a mate-
    rial error will give a sample without an error.

In many different versions of the AAM, the assurance (A) is divided
into a number of different components, such as:

-   <u>inherent assurance</u>, the measure of certainty the auditor derives
    purely from his professional judgement, his knowledge of the firm
    and of the assignment;
        (for a statistician: the subjective probability that the
        client will have made no material errors.)
-   <u>assurance from analytical review</u>, the measure of certainty that
    material errors will have been found during the performance of
    analytical review;
        (the statistician may wonder: is this the conditional probabi-
        lity of detection if it is present, or is it the joint proba-
        bility that an error is present and consequently detected.
        Further on, alike questions will be raised about the model as
        a whole, but on this point, the reader can get a preview of
        what this paper is about.)

-      <u>assurance from compliance tests</u>, the measure of certainty that
       material errors will have been found when testing on the presence
       of internal control.
       Sometimes this assurance is defined as the certainty that internal
       control itself will have found the errors, sometimes it is the
       auditor who finds them when evaluating internal control.

II.1.2. An example

In almost every application of the AAM, everybody agrees (without any
discussion) that overall assurance should equal 95%, or (what amounts
to the same thing) overall audit risk may be 5%. Assuming an auditor
specified a materiality fraction of, say, 1%, the sample size now only
depends on the level of A.

A=0   implies that β_o=0.05, so sample size n=299 (binomial),
A=0.50 β_o=0.10,   230,
A=0.90 β_o=0.50,   69,
A=0.95 β_o=1.00,   0,

and values of A > 0.95 would also render a zero sample size.

Interesting about this formula is that 'the chain can be stronger than
its strongest part': when the auditor decides A to be 50% (50% assu-
rance) and his sample has been performed to reach a β_o of 10% (90% sam-
pling assurance), the resulting overall assurance is not somewhere
between 50% and 90%, but 95%.
This example clearly shows that according to the AAM assurance from
different sources can be added, implying that a weak inherent assurance
is supposed to be compensated by a stronger sampling assurance. We will
come back to this later on.

II.2. Comments on AAM

II.2.1. Auditor's comments

First of all, auditors might object against the choice of variables in
the model.

Apart from the conviction that, at least in Dutch auditing, <u>inherent
assurance</u> is not a part of the auditor's tools and techniques, it is
difficult to see how inherent assurance can influence the range of
audit activities. At the most, it could influence the audit's objecti-
ves, not the quantity of audit activities.
At its best, <u>analytical review</u> can lead to an indication of the pre-
sence of potential errors. But it is incorrect to use information about
qualities (error rates) as if it were information about statistical
confidence (the significance level of a statistical test).

Second, the auditor may wonder whether these variables are the kinds of
quantities one can really quantify. When assessing A to be 50%, the
result of the model is in terms of assurance, not in terms of financial
units.

So, prior 'knowledge' cannot be validated by a statement about the impli-
citly assumed quality. Even a full investigation of the population will
not validate the chosen level of assurance: afterwards, a material error
was either present (0% assurance) or not present (100% assurance). It is
assumed to be a severe handicap of this model that the auditor cannot
validate his assumptions in a way that confirms his ideas about the
quality of the population subject to his audit.

II.2.2. Methodological comments

When the American Institute was still in the first stages of discussing
the notion of Audit Assurance, K.A. Smith (1972) already warned:
> 'No logical basis has been determined for setting the confidence
> level correlated with different states of internal control. The
> selection of levels to be utilized is completely arbitrary, with-
> out any theoretical basis'.

By quantifying all these forms of 'assurances' as variables that affect
(or can be supplemented to) statistical confidence levels, information
about the prevalence of error fractions is used as information about
confidence levels. In other words: the required confidence level of a
hypothesis to be tested is influenced by a prior belief about the vali-
dity of the same hypothesis.

Statisticians will not lightly support this auditors' habit. Statisti-
cians will argue that the confidence level of a statistical test must
be set before the actual test is performed, and should not be affected
by any prior idea about the trueness of the hypothesis to be tested.

The AAM, though, suggests that a weak inherent assurance can be compen-
sated by a stronger sampling assurance, or a strong inherent assurance
suffices with a weak sampling assurance. The only logical basis behind
this would be that statistical confidence is a statistical variable,
that could be transferred from 'belief in the trueness of a theory' to
its empirical validation. As if a strongly believed theory only has to
be validated by a weak statistical result, and less strongly believed
theories need more statistical support.
On the contrary: the measure a theory is believed to be true does not
affect the confidence level it is tested at, although the stronger a
theory is believed to be true, the stronger the expectation of empiri-
cal evidence will be when that belief is tested.

> (An elaboration on this statement could even trigger off a vivid
> discussion amongst statisticians, so let's take an example: every-
> body knows that personal income is the main variable that affects
> individual consumption. When, in a regression analysis on a large
> sample, the income coefficient is tested on significance, no eco-
> nometrician will use his certainty that income is the main expla-
> natory variable to raise (=to weaken) the requested significance
> level. Econometricians would even be disappointed when significan-
> ce could **only** be reached at 5%, as they expect their ('certain')
> theoretical knowledge to become confirmed more firmly.)

Statistical confidence is not a statistical variable, and an individual value used in an individual application cannot be validated afterwards. As mentioned above, even a full investigation of an individual population will not validate the chosen level of statistical confidence: afterwards, a material error appears to be either present (0% confidence) or not present (100% confidence).

II.2.3. Comment on statistical computations

Apart from a discussion about the nature of the variables in the model, there is a question of statistical independence. Amongst many others, Roberts (1978) as well as Bailey (1981) mention this question, and both tend to doubt the presence of independence. Unfortunately, neither of them draws a conclusion on the validity of the model as a whole.

In the AAM, overall assurance is defined as 1 minus the probability that neither preceding audit phases, nor subsequent statistical sampling, detects a material error. This probability is derived by multiplication of 1 minus the 'assurance A' with the probability of non-detection of a material error in the sample.

This multiplication of probabilities is only permitted when the variables referred to are statistically independent.

> **Statistical independence would imply that the probability of error-detection in a statistical sample is identical for errors that have already, and errors that have not yet been detected in preceding audit phases.**

This notion of statistical independence in fact only makes sense if the related variables are statistical variables, but we already stated that 'assurance' can not be interpreted as such.

However, even if we -just for argumentation- interpret 'assurance' as the probability of detection of a material error, the assumption of statistical independence has not yet been proved to be correct.

Therefore, as long as it is not validated, auditors should not rely on this assumption, but should stay on the safest side. When determining overall assurance from assurance and sampling assurance (sampling confidence), the auditor has to start from the most unfavorable combination of both. This is the situation in which audit sampling renders as little extra information as possible, because detection of errors in all audit phases overlap as much as possible, resulting in the detection of errors in the sample that already were detected in preceding phases. When sample size is sufficient (large enough) to reach the required overall assurance in this situation, it is always sufficient. When determining overall assurance under the most unfavorable combination of assurances, the result is as disappointing as it is predictable: in section 3 it will be shown that overall assurance is equal to the maximum of the individual assurances. This implies:

- the larger inherent assurance, or assurance from analytical re-
  view, the larger statistical assurance (and the smaller sampling
  risk) must be in order to render a sufficient sample;
- only when statistical assurance is chosen equal to the required
  overall assurance, the auditor is sure to have a sample that is
  always large enough to meet his requirements.

The more preceeding assurance obtained, the larger sample is required to
validate the auditor's judgements. This conclusion is not very at-
tractive to the auditor, but is not therefore illogical: the stronger a
theory is believed to be true, the stronger empirical validation is
necessary to strengthen that belief.

II.3. Mathematical proof and graphical illustration

II.3.1. Mathematical proof

To show that a statistically improved version of AAM gives the result
we mentioned above, we make a (2X2)-chart of possible events and their
(assumed) probabilities. (Chart 1)
In the sample, a material error is detected (regardless of whether it
was already detected by previous audit activities) with probability
$1-\beta_o$ and not detected with probability β_o, and the AAM <u>interprets</u> $1-A$
to be the probability that an error has not been found in previous
stages of the audit and A its complement, the probability that the
error has already been detected.

Overall assurance is now derived by filling in the inner part of the
chart. To reach the expression for overall assurance given in the AAM
the marginal probabilities are multiplied.
As we can see from chart 2, overall assurance, the probability that
either previous activities, or sampling, or both, will detect a mate-
rial error, is equal to 1 minus the probability that neither will find
it:

$$OA= 1- \beta_o(1-A).$$

Implicitly, by multiplying these probabilities, independence between
previous audit activities and sampling has been assumed. What will
happen if we drop this assumption?

To answer this question, we make 4 different charts out of chart 1. Re-
stricted by the marginal probabilities, we can investigate the extreme
values of the probability not to find a material error.
In each chart, a (different) corner is filled by the lowest possible
probability, zero. The rest of the chart is completed using the known
marginal probabilities. In this way, we can see between which lower-
and upper limits the value of the Overall Assurance, 1 minus the proba-
bility in the lower-right corner, lies.

sample result:	result of previous activities: error found	error **not** found	
error is detected			$1-\beta_o$
error is **not** detected			β_o
chart 1	A	1-A	1

sample result:	result of previous activities: error found	error **not** found	
error is detected	$A(1-\beta_o)$	$(1-A)(1-\beta_o)$	$1-\beta_o$
error is **not** detected	$A\beta_o$	$\beta_o(1-A)$	β_o
chart 2	A	1-A	1

Charts 3 and 4 result in an upper limit for OA. This upper limit of OA is 1 when chart 4 is accurate, that is, when $A-\beta_o > 0$, and it is $1-(\beta_o-A)= 1+(A-\beta_o)$ when $A-\beta_o < 0$. In the latter case, OA is <1. There-fore, we can conclude:

   The upper limit of OA, the maximum value that can be reached, is the minimum of 1 and $1-(\beta_o-A)$.

Charts 5 and 6 give information about the minimum value of OA. Chart 6 shows that OA = $1-\beta_o$ when $1-\beta_o-A > 0$, so when $1-\beta_o > A$, and Chart 5 shows that OA=A when $1-\beta_o-A < 0$, so when $A > 1-\beta_o$. Conclusion is:

   The lower limit of OA, the minimum value that will be reached, is the maximum of A and $1-\beta_o$.

Together:

   $\max \{A, 1-\beta_o\} \leq OA \leq \min \{1, 1-\beta_o+A\}$.

Translated for auditors: overall assurance can not be calculated, be-cause it is not known how previous audit activities affect the probabi-lity that a material error will be discovered in the sample.

Chart 3

result of previous activity:	error found	not found	
sample result: error detected	0		$1-\beta_o$
NOT detected			β_o
chart 3	A	1-A	1

result of previous activity:	error found	not found	
sample result: error detected	0	$1-\beta_o$	$1-\beta_o$
NOT detected	A	β_o-A	β_o
	A	1-A	1

Chart 4

result of previous activity:	error found	not found	
sample result: error detected			$1-\beta_o$
NOT detected		0	β_o
chart 4	A	1-A	1

result of previous activity:	error found	not found	
sample result: error detected	$A-\beta_o$	1-A	$1-\beta_o$
NOT detected	β_o	0	β_o
	A	1-A	1

Chart 5

result of previous activity:	error found	not found	
sample result: error detected		0	$1-\beta_o$
NOT detected			β_o
chart 5	A	1-A	1

result of previous activity:	error found	not found	
sample result: error detected	$1-\beta_o$	0	$1-\beta_o$
NOT detected	$A-1+\beta_o$	1-A	β_o
	A	1-A	1

Chart 6

result of previous activity:	error found	not found	
sample result: error detected			$1-\beta_o$
NOT detected	0		β_o
chart 6	A	1-A	1

result of previous activity:	error found	not found	
sample result: error detected	A	$1-\beta_o-A$	$1-\beta_o$
NOT detected	0	β_o	β_o
	A	1-A	1

There is only a minimum and a maximum value of overall assurance; the auditor might aim at maximizing the minimum value. When he assesses a value for his 'assurance', A, he can decide whether this value would already be sufficient. In that case, no sample is necessary. If not, his sample assurance should equal the value of overall assurance required:

$$1-\beta_o = OA.$$

Consequently, the higher value of A, the smaller β_o must be chosen to render a sample that provides a gain in assurance over A. In other words: the more favorable prior knowledge, the larger the sample must be, before it is of use. This conclusion is more than just a tendentious remark: it is completely coherent with our methodological point of view from section 3.2.2.

II.3.2. Graphical illustration

Users of the AAM often use this example to explain their method. In a circus, we want to prevent the trapezists from falling on the floor by stretching several rope nettings. The first netting has already been hung; this plays the part of (inherent) assurance (on internal control etc.). When we require a certain overall assurance OA (or an Audit Risk AR), how large must our second netting (the auditor's sample) be?
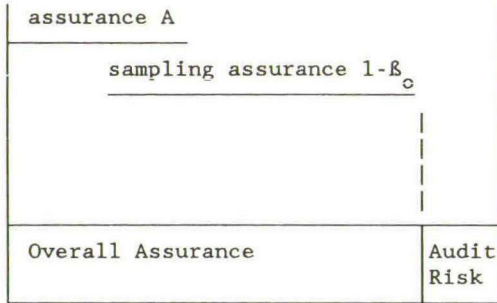
Next, there appears a drawing (figure 1) of the two nettings. As we can see, both nettings together do yield the required probability of intercepting the trapezist.
But, we wonder, why is there an overlap ? Isn't that inefficient ? Why not hang our second netting like (1) in figure 2 ? Then, a much smaller netting (= sample size) would be sufficient ! Or even, why not use a bit larger netting, like (2), and attain 100% assurance !
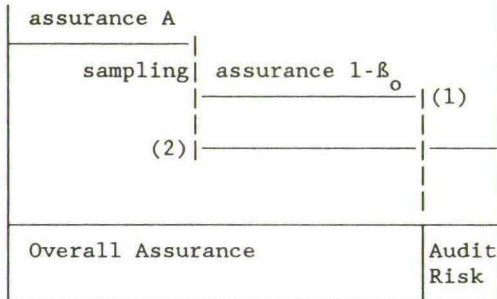
The problem is that we can choose the length of the netting (sample size), but we cannot set the measure of overlapping between both nettings, the dependency between previous audit activities and sampling. The only way to be sure that the netting is large enough even when it hangs in the worst place, is to take a netting that is as large as the overall assurance required.

Again, we come to the conclusion that dropping the independence assumption from the AAM, and assuming that A < OA, results in a sampling risk β_o equal to 1 minus overall assurance: the stronger a theory is believed to be true, the stronger empirical validation is necessary to strengthen that belief.
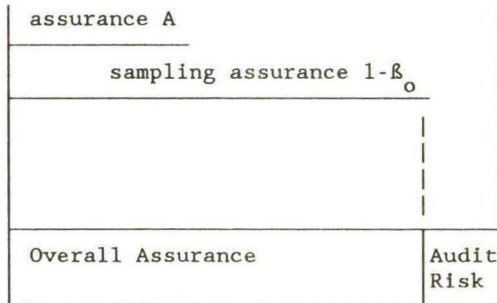
(fig. 1)

```
┌─────────────────────────────────────────────┬──────────┐
│ assurance A                                  │          │
│ ─────────                                    │          │
│        sampling assurance 1-ß                │          │
│                            o                 │          │
│        ─────────────────────────            │          │
│                              |               │          │
│                              |               │          │
│                              |               │          │
│                              |               │          │
├─────────────────────────────────────┬────────┤
│ Overall Assurance                    │ Audit  │
│                                      │ Risk   │
└──────────────────────────────────────┴────────┘
```

(fig. 2)

```
┌─────────────────────────────────────────────┬──────────┐
│ assurance A                                  │          │
│ ─────────|                                   │          │
│        sampling|  assurance 1-ß              │          │
│                |               o    |(1)     │          │
│                |───────────────────|         │          │
│                |                    |──       │          │
│            (2)|────────────────────|         │          │
│                |                    |         │          │
│                |                    |         │          │
├─────────────────────────────────────┬────────┤
│ Overall Assurance                    │ Audit  │
│                                      │ Risk   │
└──────────────────────────────────────┴────────┘
```

(fig. 3)

```
┌─────────────────────────────────────────────┬──────────┐
│ assurance A                                  │          │
│ ─────────                                    │          │
│        sampling assurance 1-ß                │          │
│                            o                 │          │
│                              |               │          │
│                              |               │          │
│                              |               │          │
│                              |               │          │
├─────────────────────────────────────┬────────┤
│ Overall Assurance                    │ Audit  │
│                                      │ Risk   │
└──────────────────────────────────────┴────────┘
```

II.4. Conclusion to part II

The AAM has been shown to be a statistically doubtful formula, contai-
ning variables that should not be in it, with numerical values that can
not be validated, and giving results that are methodologically not
valid.
And, what is even worse, many auditors claim not to use it (because
they know the model is wrong), but in spite of that, let depend the
value of $ß_o$ to be used on their subjective judgement on internal con-
trol.

That too is a mistake. Statistical sampling is like the thermostat of your heater in winter time: no matter how the weather is outside, the thermostat guarantees you that the temperature you choose will be reached in your room. When you assume it will be cold outside, you should not put the thermostat up, nor should one put the thermostat down when it is warm outside.

Of course, auditor's knowledge and experience, and the results of previous audit activities, may not get wasted when the auditor comes to his audit sample. Some variables in the AAM are good ways to quantify 'professional judgement'.
The only problem is that they do not, and therefore may not, affect the confidence level used to test on a specific error fraction. They are all factors that should influence the distribution of the error fraction itself.

In a Bayesian model, (e.g. Kriens, 1963, Veenstra and van Batenburg, 1989, van Batenburg and Kriens, 1989) the same factors can be incorporated without methodological drawbacks, and in a way that the auditor can validate his professional judgement in monetary dimensions.

III Bayesian Discovery Sampling, a better method to utilize the audi-
     tor's 'professional judgement' in sampling

Part three is organized as follows. In section 1, the notion of Baye-
sian statistics is explained, in order to show the difference between
'classical' probabilities and Bayesian probabilities. As an introduc-
tion, section 2 presents a naive model of Bayesian audit sampling. In
section 3 the way is made for a less naive model, by showing the rela-
tion between interval estimation and Bayesian inference. Section 4 pre-
sents our model of Bayesian Discovery Sampling, and section 5 is about
the practical application in Touche Ross Nederland audits. Finally,
section 6 concludes.

III.1. Bayesian inference

Reverend Thomas Bayes (1702-1761), in his search for methods to design
experiments that proved Newton's ideas about the laws of nature (see K.
Pearson, 1978), gave name to a whole new way of looking at probabili-
ties. Bayes showed how probabilities can be (re-)defined using both
prior knowledge about the event itself and empirical evidence from
sample results.

Beginning students in statistics are often confronted with the standard
Bayes-problem: two vases, labeled 1 and 2, contain, in different but
known proportions, red and white chips. First a lot is drawn in order
to decide randomly which vase is used, and from that vase one chip is
drawn at random. The probability distribution of the color of the chip
is now dependent on the label of the vase. Bayes showed how - vice
versa - the color of the drawn chip affects the probability that vase
1, or vase 2, has been chosen.

Say, for example, that vase 1 contains 6 red and 4 white chips, and
vase 2 contains 3 red and 7 white chips, and vases are drawn each with
50% probability. Now if the drawn chip is red, according to Bayes'
theorem, there is a posterior probability of 2/3 that vase 1 has been
chosen, and 1/3 that it was vase 2.

The modern version of Bayesian inference says that the probability
distribution of possible events (the colors of the chips) depends on
unknown parameters (the labels of the vases). These parameters themsel-
ves have (subjective) probabilities, which are defined in a probability
distribution (in the example 50% for each vase). This distribution is
the prior distribution of the parameters, quantifying the probabilities
of the possible parameter values prior to empirical investigation.

When the experiment is completed, and empirical results have become known, we can formulate a posterior distribution, 'updating' the probabilities of these parameters in the light of the empirical results.

Translated to auditing, the same example can be used referring to an auditor who wants to evaluate a population. He lays down a standard for what is 'good' and what is 'bad' (the labels of the vases) and specifies his subjective prior probabilities of 'good' and 'bad'. (This prior distribution in general will of course not be 50% for each alternative.) The conditional distribution of the possible sample results, that is the number of errors in the sample (the colors of the chips) can be derived for both the 'good' and the 'bad' population, respectively.

After the sample has been drawn and audited, we can, retrospectively, calculate the posterior probabilities of a 'good' or a 'bad' population, given the objective sample results and taking into account the original subjective ideas about the probabilities of a 'good' and of a 'bad' population.

In this way, the auditor evaluates the population, not only by the objective sample results, but also by his prior professional judgement.

III.2. A naive Bayesian model

Suppose an auditor knows a priori that the population to be audited is either 'good' (p, the population error fraction, is 0), or 'bad' (it contains a certain fraction of, p_1). Furthermore, the auditor assigns a prior probability of 1-q to p=0 and, thus, q to $p=p_1$. Thinking in Bayesian terms, we can say that without any additional information (e.g. sample results) the posterior probabilities are equal to the prior probabilities, so also 1-q for p=0 and q for $p=p_1$.

When a sample of size n is audited, every 'good' item will increase the posterior probability of a 'good' population, whereas a 'bad' item (an error) decreases this probability (and increases the posterior probability of a 'bad' population).

It is not that difficult to calculate the sample size n, that, with n 'good' items and zero errors, increases the posterior probability of a 'good' population to a level that is sufficient for the auditor to base his (positive) final judgement upon.

Chart 7 gives the prior probabilities, and in chart 8 these have been combined with the conditional probabilities of the sample results. Chart 8 is derived from the fact that if p=0, the probability of a perfect sample is 1, and if $p=p_1$, this probability is $(1-p_1)^n$. From the first row of chart 8, we can calculate the posterior probability of $p=p_1$, given a perfect sample ($\underline{k}=0$). We can see that the fact that $\underline{k}=0$ has decreased the probability for $p=p_1$ from q to:

$$P(\ \underline{p}=p_1\ |\underline{k}=0\ )=\frac{q(1-p_1)^n}{(1-q)\ +\ q(1-p_1)^n}$$

We can calculate the minimum sample size n for which this posterior probability of wrongly accepting the 'bad' population, given $\underline{k}=0$, is less than, or equal to $ß_o$. From that calculation follows:

sample result:	prior knowledge: p=0 p=p_1	
no errors detected		
1 or more detected		
chart 7	1-q q	1

sample result:	prior knowledge: p=0 p=p_1	
no errors detected	$(1-q)*1=1-q$ $q(1-p_1)^n$	
1 or more detected	$(1-q)*0=0$ $q[1-(1-p_1)^n]$	
chart 8	1-q q	1

n= $\log\ \{(ß_o(1-q))/((1-ß_o)q)\}/\log\ (1-p_1)$.

Compared to the sample size in 'classical' Discovery Sampling (n*), we expect the above sample size to be smaller as long as our prior knowledge is in favor of the population being 'good', so as long as q is less than 50%.

(Exact calculation gives: n<n* iff q< $1/(2-ß_o)$, which is a little more than 50%, but the rationale for this negligible difference is beyond the purpose of this paper.)

Before going in to the naivety of this model, we give one numerical example.
Discovery Sampling with $ß_o$=0.05 and p_1= 1% gives a classical sample of 299. Suppose the auditor is 75% sure of a perfect population (and thus assigns a 25% probability to the population containing a material error). If he aims at a posterior probability of 0.05 (corresponding to $ß_o$=0.05) for the population being 'bad', sample size reduces to 184 !

In a graph (Figure 3), we can illustrate this method as follows. Before
sampling, there is a prior probability distribution with density $1-q$ on
$p=0$. When the sample consists of n 'good' and 0 'bad' items, enough
probability has been moved from $p=p_1$ to $p=0$, to make the posterior
density of $p=0$ equal to $1-\beta_0$.

(Figure 3)
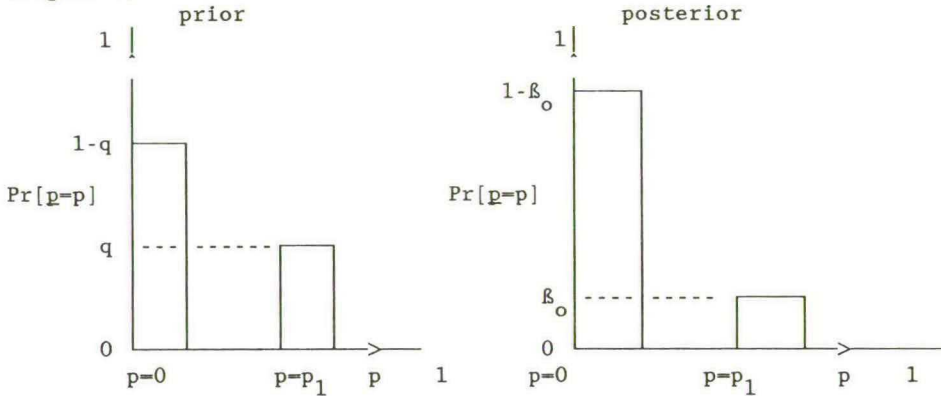
(Figure 4)
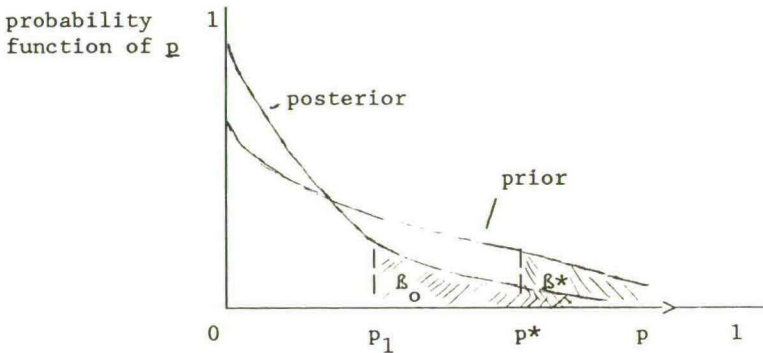
Of course, this model is too naive to use in auditing.
The population error fraction is not either 0 or p_1, but has a value in
a range that is theoretically bounded by 0 and 100%. In our real model,
we will use the assumption that the auditor 'knows' (with a specific
certainty) that this range will not be 0-100%, but, say, 0-75%.
But still, as will be explained in section 6, that model works with the
same basic idea (like in Moors, 1983) that the auditor specifies his
prior knowledge, and that a sample in which no errors occur yields a
particular posterior probability of wrongly accepting the (bad) popula-
tion. Sample size is, just as above, calculated from the restriction
the auditor imposes on this posterior probability.

III.3. On interval estimation and Bayesian reasoning

As we saw in section 2, Discovery Sampling is based on the calculation
of the upper limit of a (in our case) 95% confidence interval for the
population error fraction p, when no errors have been found in the
sample. The size of this sample must be sufficient to make this upper
limit not to exceed the designated materiality fraction.

When calculating such an interval, the statistician will start by for-
mulating the possible values of p, and consequently reduces the width
of that interval on the basis of the empirical results.
So, before a sample is taken, the possible values of p are 0-100%. Any
additional sample outcome result will result in a somewhat smaller
interval. Furthermore, a 'good' result shifts the interval towards p=0,
and a 'bad' result shifts it away from p=0. Sampling can be stopped
when the upper limit has descended from 100% to p_1.

The number of good items it takes to bring the upper limit down to p_1
does not only depend on p_1, but also on the location of this upper
limit at the start of this procedure. Is it really true that without
sampling the upper limit is theoretically equal to 100%?. In classical
statistical theory, yes, but supported by Bayesian statistics we can
start from a subjectively chosen upper limit, resulting from professio-
nal judgement and prior knowledge.

The model described in the next section therefore starts by formulating
that subjectively chosen upper limit. From that point, the sample size
is calculated to derive the upper limit aimed at by the auditor. In
figure 4 is shown what will be mathematically formulated in Section 4.

III.4. Bayesian Discovery Sampling

(1)   As prior probability function for the unknown error fraction in
       the population we choose:

       $Pr(p) = s(1-p)^{s-1}$ for $0 \leq p \leq 1$ and $s > 0$.

       This very simple prior has only one parameter, s, so it takes only
       one result to find a value for it. Furthermore, it has its mode in
       p=0, which is consistent with the auditor's expectation that the
       population contains very few errors.

The parameter s is chosen in accordance to the evaluation of last year's audit sample. We suppose that in the previous year, discovery sampling has been performed with parameters $\beta*$ and $p*$ and that no errors have been found. This implies that the upper limit of the $100(1-\beta*)\%$ confidence interval for p was $p*$. (If errors have been found, a value of $p*$ can be calculated that is larger than the materiality fraction in that audit, but it does not change our model.) So, s can be set at that value that results in a probability $\beta*$ for $\underline{p}$ exceeding this upper limit $p*$:

(Strictly speaking, this probability might even be taken equal to $\beta*(1-p*)$, but for simplicity we have ignored this subtlety.)

$$P\{ \underline{p} > p* \} = \beta*$$

$$P\{ \underline{p} > p* \} = \int_{p*}^{1} \Pr(p) \, dp = \int_{p*}^{1} s(1-p)^{s-1} \, dp = [-(1-p)^{s}]_{p*}^{1} = (1-p*)^{s};$$

from $(1-p*)^{s} = \beta*$ it follows that $s = \log \beta*/\log(1-p*)$.

(2) The probability of an errorless sample of size n from a population of size N with error fraction p can be approximated by:

$$L(\underline{k} = 0 \mid p, n, N) = (1-p)^{n}.$$

Of course, we could have used the fact that the sample is taken without replacement. On the other hand, as seen in section 2, it is quite common to disregard this fact. We even intentionally chose our prior to have a general form that is mathematically easy to combine with this sample likelihood. (In statistical handbooks, e.g. Zellner, 1971, the term 'natural conjugate prior', is used for a general form that simplifies the calculation of the posterior.)

(3) The posterior probability function for p is derived from (1) and (2). Mathematically it is a bit more difficult but actually not different from the way the posterior probability was derived in section 2:

$$Po(p \mid k=0, n, N) = (n+s)(1-p)^{n+s-1} \text{ for } 0 \leq p \leq 1, \ n+s>0.$$

(4) This posterior function has to meet the auditor's requirements for discovery sampling in this year. That means, that the parameter $(n+s)$ has to reconcile the information that would have resulted from the upper limit of a $100(1-\beta_o)\%$ confidence interval, and that this upper limit should equal p_1. This means (apart from the subtlety just mentioned when discussing the prior):

$$P\{ \underline{p} > p_1 \} = \beta_o$$

$$P(\underline{p}>p_1) = \int_{p_1}^{1} Po(p) \, dp = \int_{p_1}^{1} (n+s)(1-p)^{n+s-1} \, dp = [-(1-p)^{n+s}]_{p_1}^{1} = (1-p_1)^{n+s};$$

from $(1-p_1)^{n+s} = \beta_o$ it follows that n+s= log $\beta_o/\log(1-p_1)$.

(5) Combining the expressions for s in (1) and n+s in (4), we get a sample size n_B that is sufficient for Bayesian Discovery Sampling with parameters β_o and p_1, based on prior knowledge incorporated in $\beta*$ and $p*$:

n_B = log $\beta_o/\log(1-p_1)$ - log $\beta*/\log(1-p*)$.

(6) In practice, it will be rather unrealistic for the auditor to state that last year's audit sample evaluation is fully giving the right prior information for this year's prior probability function. Therefore, we incorporate a weight function f:

$n = f * n_B + (1-f) * n_C$

In this function, n_C is the classically determined sample size and f is the weight $(0 \leq f \leq 1)$ the auditor gives to his prior information, that is the extent to which he 'dares' to lean on his subjective prior knowledge. The size of the sample to perform is thus a weighted average between the Bayesian sample size n_B and the classically determined sample size n_C (which equals log $\beta_o/\log(1-p_1)$).

A little substitution gives:

n = log $\beta_o/\log(1-p_1)$ - f. log $\beta*/\log(1-p*)$

It is easy to see that an auditor who does not want to use his knowledge based on last year's sample and sets f at 0, gets n = n_C. On the other hand, when an auditor completely leans on his prior knowledge (that is, on last year's errorless sample), and his audit parameters have not changed since last year (so β_o =$\beta*$ and p_1= $p*$), the above calculations will result in a zero sample size.
The latter is not a problem from a statistical point of view, but it may be undesirable from the point of view of an auditor or an auditor's firm.

Before we give some details, in the next section, as to how f is chosen in Touche Ross Nederland practice, we will present some numerical examples. Suppose that last year, an auditor has audited a sample of 59 items when performing discovery sampling with $\beta*$=0.05 and $p*$=5%, and that no errors were found. In this year he once again chooses β_o =0.05 and p_1=5%. Using classical theory, a new sample of 59 would be required.

However, the auditor uses his prior knowledge (and everything else he is used to do when applying the Audit Assurance Model) and decides that f is 70%. He can now choose between:

- performing a sample of 19, which is sufficient for β_o =0.05 and p_1=5%, because n = 59 - 0.7 * 59 = 19;

- performing a sample of 59, which is sufficient for β_o=0.05 and p_1=3%, because n= 99 - 0.7 * 59 = 59.

(Actually, these calculations lead to 18 and 58, but 0.7 times 59 is rounded down, and the final result is rounded up, just for precaution. Also in the examples coming ahead, we have not always been consistent in our rounding-offs and rounding-ups. We have been consistent in precaution.)

Another numerical example, which will be referred to in the next section, is an auditor who decides to use this year β_o=0.05 and p_1=0.5% (classical sample size 598), while last years' sample was 299 with $\beta*$=0.05 and $p*$=1%. (Let us not go into reasons why the auditor suddenly halves his materiality, these figures are just handy to explain the model.)
Assume the auditor has taken f=40%. The sample size will be:

p_1 =0.5% β_o =0.05	f=0.40, p* =1% $\beta*$= 0.05	
n= 598 - 0.4 * 299	= 598 - 119	= 479

III.5. Practical application and implementation

III.5.1. Assigning a value to f

After accepting a first report explaining the statistical method, the Board of Governors of Touche Ross Nederland has granted a budget to a committee of auditors with the task to design a method by which the auditor, in a specific audit, can assign a specific value to the factor f.

Because of the preliminary status of their results, we will not go into detail on this subject. Headline of their conclusion will be in conformity with the results of the Touche Ross Nederland audit process UNICON and its computer assisted audit planning system COCON.

In the audit approach UNICON, three phases can be distinguished:

I   Audit planning, leading to an evaluation of internal controls, a design of the audit approach and an audit program;
II   Interim-audit, an analytical review to decide on a choice between a compliance approach and a substantive approach to the audit;
III   Financial statements audit, consisting of substantive testing, balance sheet review and evaluation of audit results.

In the expert system COCON, for each audit cycle potential errors have been adapted in a database. All possible measures of internal control are specified, and the auditor can evaluate their design, their presence and their functioning. Not every measure of internal control is necessary, but the combination of individual measures should be sufficient to give the auditor an opinion on the reliance on internal control.

In COCON is a database with audit expert's opinions on how to weigh these measures, which we could call (internal) Control Evaluation Model-scores (CEM-scores).

These CEM-scores from COCON will also be the mainstay for the factor f that an auditor can assign to his specific application of Bayesian Discovery Sampling.

III.5.2. Validation of one specific application for the auditor

When performing an audit and deciding on the factor f, the auditor needs an instrument to validate the a priori added subjective informa-tion, which will reduce the necessary sample size. Of course, it is impossible to validate the notion of 'weight given to prior knowledge' or 'weight given to last year's audit sampling results'. What can be done, is validating the consequences of a specific choice of the factor f. To show this, we use the second example in section 6:

p_1 =0.5% ß_o =0.05	f=0.40, p* =1% ß*= 0.05	
n= 598	- 0.4 * 299	= 598 - 119 = 479

(The reader may assign a monetary value to the audited population, if that clears his view on the practical consequences. We did not, because every result presented here is mathematically independent of population size.)

As we can see, the consequence of choosing f=0.40 is that the auditor has implicitly decided that a sample of 119 items is errorless, without having actually audited these items this year. In other words, last year's audit sample, internal control (and all the other elements that built up 'assurance' when the Audit Assurance Model would have been used) have given the auditor a 'professional judgement' that makes him almost sure (95%) that the error fraction in this year's population will not exceed 2.5%.

(Remember the rule of thumb in section I.2.: an errorless sample of 120 items is sufficient for a 95%-upper limit of 2.5%, as 120*0.025=3. Exact calculation gives 119 instead of 120.)

In this case, the auditor, applying some statistical calculations, can validate his choice of the factor f by asking himself:

'if I want to know (with 95% certainty) whether the error fraction is below 0.5%, may I lean on my prior professional judgement that it is (with 95% certainty) below 2.5% ?'

Of course, a specific f will not always result in the same implicitly chosen upper limit for p: this upper limit not only depends on f, but also on last year's sample size.

III.5.3. Validation of various applications for the auditor's firm

Apart from the individual auditor, other parties are concerned in the validation of the use of prior information. One of those is the auditor's firm that carries out mutual quality control on the performances of individual auditors. In retrospect, the method can be validated by the following reasoning. In the example already mentioned, the auditor has taken a sample of 479 items, and evaluated it as if it were 598 items. If one wants to know whether this decision has been made on justified grounds, the obvious thing one can do is to audit as yet those lacking 119 items!

In fact, there are two ways of reasoning, both leading to the same result. First one can state that the prior assumption $p \leq 2.5\%$ has to be tested, for which an errorless sample of 119 items is sufficient. The second manner is to state that the overall evaluation $p \leq 0.5\%$ must be investigated, for which an errorless sample of 479 is not, but an additional errorless sample of 119 again is sufficient.

In order to make such a validation possible, a sample of 598 items will be drawn, of which 479 are randomly selected to be audited. The remaining 119 will also be audited in case the auditor decides not to use his prior information (the auditor decides to lower f to 0), or when this is required for the purpose of validation.

III.6. Conclusion to part three

In this paper, we offered a first look on the results of a project that started about 3 years ago, based on a paper that was written 25 years ago (Kriens, 1963). We ourselves are quite sure that it will take at least another 3 years before our auditors can apply this method, but less than 25 years before it is really optimally applied by every auditor in Touche Ross Nederland.

Our goal was to give an alternative for the Audit Assurance Model, about which almost every auditor knows 'it is not perfect', but only few auditors realize how misleading it is. The Bayesian approach is there as an alternative, and maybe its best characteristic is that auditors can use their habitual methods, the audit program they were used to in the Audit Assurance approach, in this Bayesian alternative. The only difference, some stubborn auditors might say, between the 'old' and the 'new' approach is that it is based on (in-)accuracies, and not on confidence levels: it implies only a change in statistics, not in auditing.

Statisticians will answer this is a slight - but most significant- understatement.

References

BAILEY, A.D.(1981) **Statistical Auditing.** New York, HBJ.

BATENBURG, P.C. van and J.KRIENS (1989) "Bayesian discovery sampling: a simple model for Bayesian inference in auditing". **The Statistician**, 38: p. 227-233.

BLYTH, C.R.(1986) "Approximate Binomial Confidence Limits." **Journal of the American Statistical Association** 81: p. 843-855.

KRIENS, J. (1963) "De methoden van de Wolff en van Heerden voor het nemen van aselecte steekproeven bij accountantscontroles". **Statistica Neerlandica** 17: p. 215-231.
    De Wolff's and Van Heerden's methods for random sampling in auditing, in Dutch.

MOORS, J.J.A.(1983) "Bayes' estimation in sampling for auditing." **The Statistician** 34: p. 281-288.

PEARSON, E.S.(ed.)(1978) **The history of statistics in the 17th and 18th centuries: lectures by Karl Pearson.** London, Griffin & Co.

ROBERTS, D.M.(1978) **Statistical Auditing.** New York, American Institute of Certified Public Accountants.

SMITH, K.A. (1972) "The Relationship of Internal Control Evaluation and Audit Sample Size". **The Accounting Review**: p. 260-269.

VEENSTRA, R.H. and P.C. van BATENBURG (1989, 1990) "Een doorbraak in steekproeftoepassingen door Bayesiaanse statistiek." **De Accountant** 11(1989): p. 561-564 and 1 (1990): p. 18-21.
    "A breakthrough in statistical applications by Bayesian statistics", in Dutch. Published in the magazine of the Nederlands Instituut van Registeraccountants (NIVRA).

ZELLNER, A.(1971) **An introduction to Bayesian inference in econometrics.** New York, Wiley.

IN 1990 REEDS VERSCHENEN

419 Bertrand Melenberg, Rob Alessie
     A method to construct moments in the multi-good life cycle consumption model

420 J. Kriens
     On the differentiability of the set of efficient (μ, σ^2) combinations in the Markowitz portfolio selection method

421 Steffen Jørgensen, Peter M. Kort
     Optimal dynamic investment policies under concave-convex adjustment costs

422 J.P.C. Blanc
     Cyclic polling systems: limited service versus Bernoulli schedules

423 M.H.C. Paardekooper
     Parallel normreducing transformations for the algebraic eigenvalue problem

424 Hans Gremmen
     On the political (ir)relevance of classical customs union theory

425 Ed Nijssen
     Marketingstrategie in Machtsperspectief

426 Jack P.C. Kleijnen
     Regression Metamodels for Simulation with Common Random Numbers: Comparison of Techniques

427 Harry H. Tigelaar
     The correlation structure of stationary bilinear processes

428 Drs. C.H. Veld en Drs. A.H.F. Verboven
     De waardering van aandelenwarrants en langlopende call-opties

429 Theo van de Klundert en Anton B. van Schaik
     Liquidity Constraints and the Keynesian Corridor

430 Gert Nieuwenhuis
     Central limit theorems for sequences with m(n)-dependent main part

431 Hans J. Gremmen
     Macro-Economic Implications of Profit Optimizing Investment Behaviour

432 J.M. Schumacher
     System-Theoretic Trends in Econometrics

433 Peter M. Kort, Paul M.J.J. van Loon, Mikulás Luptacik
     Optimal Dynamic Environmental Policies of a Profit Maximizing Firm

434 Raymond Gradus
     Optimal Dynamic Profit Taxation: The Derivation of Feedback Stackelberg Equilibria

IN 1991 REEDS VERSCHENEN

466   Prof.Dr. Th.C.M.J. van de Klundert - Prof.Dr. A.B.T.M. van Schaik
      Economische groei in Nederland in een internationaal perspectief

467   Dr. Sylvester C.W. Eijffinger
      The convergence of monetary policy - Germany and France as an example

468   E. Nijssen
      Strategisch gedrag, planning en prestatie. Een inductieve studie
      binnen de computerbranche

469   Anne van den Nouweland, Peter Borm, Guillermo Owen and Stef Tijs
      Cost allocation and communication

470   Drs. J. Grazell en Drs. C.H. Veld
      Motieven voor de uitgifte van converteerbare obligatieleningen en
      warrant-obligatieleningen: een agency-theoretische benadering