Belastingdienst

**Date**
17 February 2017

**Authors**
Emma Gottesman (lead data scientist)
Roel Niessen (data scientist)

**From**
Data and Analytics department
The Netherlands Tax Administration

**To**
IOTA GPG Project 'Applying Data and Analytics in Tax Administrations'



memo

# The importance of random samples

## Introduction

Random samples are important to the tax administration for several reasons. Benefits of the random sample usually do not appear immediately, but take some time to materialize. They are of vital importance in the long run.

This document describes the importance of random samples for the tax administration in general. It also describes the way random samples are used in the Dutch tax administration. These document is built up along the following ten sections.

1. Why random samples are important
2. Background information on random samples
3. Required size of random samples
4. Baseline and providing value
5. Tax gap and similar calculations
6. Mitigating against bias
7. Helping to test models
8. One year is not enough
9. Quality samples or re-audits
10. Random samples at the Dutch Tax Administration

## 1. Why random samples are important

Random samples are of interest to the tax administration for various reasons. These reasons are mentioned here briefly, a more thorough explanation follows later in this document.

- Random samples enable the building of risk models for tax types and processes where no models exist yet – in particular where no audit process is currently in place.

- Random samples ensure the quality of our rules and models by expanding the view of the tax administration beyond known fraud and error patterns.
- Random samples help prevent the rules and models from becoming outdated.
- Random samples help measure the tax gap by providing an unbiased and representative view of the complete population.
- Random samples make sure that each tax payer has a chance of getting caught.
- Random samples provide a baseline to measure the performance of as-is and new processes against.

## 2. Background information on random samples

A *sample* is a set of companies, individuals or tax returns selected for examination or audit. These samples can be selected in many ways, for instance every tax return requesting a repayment over a certain amount, or every return where a certain allowance is claimed.

A *random* sample is where the company, individual or return is selected at random in some way.

This can be completely at random using a random number generator: for instance a random number between 0 and 1 is generated every time a tax return is received, and if that number is greater 0.98 then that return is selected for audit. This would randomly select 2% of the tax returns.

Or we can use *stratified random sampling* where some other constraint is applied. As an example we may be particularly interested in getting extra returns that exhibit a certain characteristic, such as owning shares. So we might select a return with share ownership if the randomly generated number is greater than 0.80 whilst leaving the threshold at 0.98 for the rest of the tax returns. This would have the effect of selecting proportionally more returns where the taxpayer holds shares than we'd get otherwise, whilst still retaining the property of randomness – we'd get 20% of the tax returns with share ownership, and 2% of the rest of the tax returns, with both sets still being randomly selected.

The reason that a *random* sample (stratified or not) is so important is that statistically the properties of a randomly selected sample mirror the population from which it was taken. This is not true if we select using business rules or risk models. We'll discuss this more later in this document.

**Business rules**
Rules for selecting returns for audit are based on business, tax and fiscal expert knowledge. They select for particular known instances of misdeclaration, missing information or other errors, such as 'You cannot claim Allowance A in Situation B'.

They can also be used to categorise returns into specific types. An example is the rule that identifies a tax return as being 'declared by a company'.

**Analytical models**
The analytical models used in the tax administration, work by using statistics to identify patterns in the input data that are associated with the outcome we are trying to identify. Such as patterns in a taxpayer's income tax data that are associated with under-declaration of income tax, or patterns in a company's data and VAT return that are associated with an incorrect claim for a refund. Typically these patterns are more subtle and less clear-cut than the instances selected for using rules.

**Example: a random sample in income tax**
Results from the income tax (private persons only) audits for 2013 are shown in Table 1. Here 'random sample' means the results from that year's random sample and 'all other audits' are the results from tax year 2013 from both manual audits and automatic corrections selected using business rules.

**Table 1: results from the income tax audits for tax year 2013 ***

| | random sample | | all other audits | |
|---|---|---|---|---|
| | % | mean value[1] | % | mean value |
| **over-declared** | 5% | -€100 | 15% | -€1500 |
| **correct** | 85% | | 50% | |
| **under-declared** | 10% | €500 | 35% | €1500 |

These results will be used in the rest of this document to provide concrete examples to make the points presented clearer. Any argument made using this data will apply to parallel situations in VAT, company tax or any tax.

[1] The average, or mean, value is quoted here to make the example and the discussions simple. In general it is not a robust measure as the distribution of value is highly skewed: most of the corrections are for small or medium values, but there are a very small number of very high values that weight the mean towards the high end. A better measure is the median, or middle value, which means that half the values are above that point and half are below. For the random sample the median for under-declarations is €250. For all other audits the median for under-declarations is €750.

## 3. Required size of random samples

For random samples to work effectively, they need to have a certain minimum size. For example, many error and fraud patterns are relatively rare, so one would like to audit as many randomly selected returns as possible, in order to catch as many different patterns as possible. In practice, audit capacity is a scarce good and is also required for other purposes. In choosing the number of random samples, one should try to balance these interests.

In Section 10 we give more details of the random sample sizes used in the Dutch Tax Administration.

## 4. Baselines and providing value

A random sample gives a clear baseline to compare processes, business rules and models against. Without a random sample we cannot tell if we are actually providing value. Note that this applies to as-is processes as well as new ones. The definition of value can be different from project to project but the principle remains the same.

Looking at the examples given in Table 1 we can see that the income tax selection process ('all other audits') is bringing in both more returns that need correcting, and is selecting returns with a higher average value, than if we selected the audit population randomly, so it is providing value.

It should be noted that it is perfectly possible to design a rule or build a risk model that performs worse than selecting at random, although one assumes that one would not do so on purpose. The only way to know that this has happened is to have a random sample to compare against.

If we had an estimate of the cost of the risk based selection versus the cost of selecting randomly, we could tell conclusively if the selection process is worth doing in a purely monetary sense. If we did not have a random sample it would be extremely difficult to make that comparison.

So a random sample gives us a baseline to compare any current or future process against and hence calculate benefit.

## 5. Tax gap and similar calculations

If we want to know how much tax is *not* being declared for the country as a whole, or for a specific population or specific tax, then we can do so by looking at the results in an audited random sample.

Let's look again at the example in Table 1. A naïve extrapolation of the results from the audits selected by business rules and risks models ('all other audits') to the approximate total number of income tax returns submitted each year gives us the values in Table 2.

**Table 2: results from the income tax audits selected by rules and model scaled to approximate total population of 10 million ***

| | all other audits | | extrapolated to 10mln | |
|---|---|---|---|---|
| | % | mean value | # | value |
| **over-declared** | 15% | -€1500 | 1.5mln | -€2.3bln |
| **correct** | 50% | | 5mln | |
| **under-declared** | 35% | €1500 | 3.5mln | €5.3bln |
| **total balance** | | | 10mln | €3bln |

Which seems to imply that there are 1.5mln people in the Netherlands over-declaring their income tax and 3.5mln under-declaring, leading to a 'tax gap' for income tax for private persons of €3bln.

But this is a highly inaccurate picture.

In order to understand what the whole population is doing we cannot extrapolate from a set of audits selected precisely because they are riskier than average. They are, by definition, not representative.

A truer picture is shown in Table 3, where we extrapolate from the random sample.

**Table 3: results from the randomly selected income tax audits scaled to approximate total population of 10 million ***

| | random sample | | extrapolated to 10mln | |
|---|---|---|---|---|
| | % | mean value | # | value |
| **over-declared** | 5% | -€100 | 0.5mln | -€100mln |
| **correct** | 85% | | 8.5mln | |
| **under-declared** | 10% | €500 | 1mln | €500mln |
| **total balance** | | | 10mln | €400mln |

The statistics of a *randomly selected* sample mirror the population from which it was taken, and so it is *representative* of that population[2].

[2] This principle is used in opinion polls where the poll designer will go to considerable lengths to ensure the randomness that allows their selection to be representative. A common reason for these polls to be inaccurate is that sufficient randomness has not been achieved so the poll results are biased and hence unrepresentative. Example: telephone polls rely on a phone number being publicly available, on the person's ability to understand the questions being asked and on their willingness to answer the questions. So these polls would miss people who do not have phones, or who do not speak Dutch, or who do not

This means we can be reasonably certain that the true rate of over or under-declarations in the population of ~10mln income tax declarations for private persons – which we cannot definitively measure as it is simply not possible to audit every one of the 10 million returns - is the same rate as that measured in the random sample.

This tells us that the 'tax gap' for income tax for private persons is actually only €0.4bln, *not* €3bln*.

Similar calculations can be done on the random sample to estimate how many people with children are claiming child-based tax credits. Or a random sample of VAT returns can tell us which types of businesses are more likely to claim incorrect refunds.

Analysis of individuals or businesses selected by business rules or statistical models cannot provide accurate descriptions of a population, or accurate measure of how a population is behaving. Only a random sample can provide that unbiased, representative view.

## 6. Mitigating against bias

In order to build risk models, which find patterns in the taxpayers' data that lead to incorrect tax returns, we need tax accounts that have been audited, so we know if there was a correction or not.

The majority of such accounts were selected for audit by business rules (or risk models) and hence contain patterns of misdeclaration that we already know about, or that existed in previous years' data. See Section 7 'One year is not enough' for more on the second point.

This means that new or undiscovered patterns that lead to misdeclaration cannot be found, unless it's by accident in a return that was selected for some other reason. This of course does happen and it is one of our ways of finding these new patterns.

Adding audit results from a random sample into the data used to build the model on allows the model to discover previously unknown or genuinely new patterns or instances of misdeclaration.

In fact, in a purely theoretical sense it is considered best practice to build risk models *only* on random samples[3]. However, it is not possible or sensible to

wish to answer questions from pollsters. So these polls are not representative of the voting population as a whole.

[3] The IRS in the USA builds some of its models on purely random samples.

do so in our case because many of the patterns or instances we need are rare enough that very few, or even no, examples would occur in our relatively small random samples.

For instance the income tax random sample is approximately 21.000 audits a year and the correction rate[4] in this sample is 10%, leading to only 2.600 corrected returns. The tax declaration form itself is over 900 fields long and covers a very wide range of topics, so there are many more ways of making mistakes that lead to an incorrect declaration than we could find in only 2.600 cases. Increasing the size of the random sample would help, but the number of cases we would need to randomly select to give a chance of complete coverage would be prohibitively expensive and a waste of resources.

Instead we use a practical balance: audits selected by rules and models so that we have plenty of examples of misdeclaration and a percentage of the audits randomly selected to mitigate against the bias and to increase the change of discovering new patterns.

## 7. Helping to test models

When a model is built we need to test it in some way before it is deployed. The best way to do this is to score a random sample with the model and then compare the model's prediction with the result of the actual audits. This allows us to make a reliable prediction of what the model will do in use, and so reduce the risk of implementing a poor model, whether in pilot or in full roll-out. The same approach can also be used to test business rules.

## 8. One year is not enough

The points above show that we need random samples for each of the business processes we work with. But why do we need to keep repeating the sample? Why isn't once enough? Three reasons: : macro and micro economic change, and changes in the tax system itself.

Economic change will cause changes in the way individuals and businesses behave financially. A company's books will look very different during a recession than they would at the high point of the economic cycle. Similar changes will occur in individuals financial behaviour as the economy changes. Even subtle shifts, such as low inflation, will change taxpayers financial data over the years.

Taxpayers will also change their behaviour as a result of a model (or a set of rules). This particularly applies to active fraud where social networks of

[4] This is the rate of over- and under-declarations combined.

offenders will quickly spread information as to what does and doesn't get flagged by a model or rules. This means that any system designed to catch misuse or error needs to keep up-to-date on new patterns emerging in the data.

Additionally the tax system itself is in a state of constant flux. Thresholds, tax rates and the amount of allowances change from year to year. These changes affect the patterns in the tax data.

All of these mean that we need to keep updating, or we risk our models and rules selecting for patterns of behaviour that no longer exist and becoming ineffective.

## 9. Quality samples or re-audits

This is a particular type of sample where what is being investigated is not the return itself, or taxpayer behaviour, but rather the quality of the original audit. These audited returns can be selected using a rule, randomly or a combination of both methods. As an example 5% of the audited returns from a particular office could be randomly selected. Here using some form of randomness in the selection method can help ensure that a particular situation, office or auditor is not unfairly chosen.

## 10. Random samples at the Dutch Tax Administration

We now focus on the requirements of the random samples from the perspective of analytics on oversight. In essence oversight is performed in two ways:

- Desk audit: a check on an individual return, carried out in tax administration offices, and where appropriate involving written or other communication with the taxpayer. If further investigation is required then a field audit can be triggered.
- Field audit: where the auditor travels to the company and examines the company's documents. This can involve examining supporting documentation for one or more returns and one or more taxes. These audits are relatively expensive to conduct.

The current predictive models in the Dutch Tax Administration are used to improve selection for desk audit, and so need to be built on data that comes from the desk audit business process. As an example: a model built to select returns for VAT desk audits needs to be built on data that can be connected to a single VAT return, and since no model is 100% accurate, on data that will allow the desk auditor to identify that there is an issue.

For small and medium enterprises (SME), much of the oversight actions are aimed to assess tax payers integrally, that is, over several returns for several tax types at the same time, using a field audit. The Dutch Tax Administration do not currently have models that assess companies in this fashion, but development of these is planned.

Random samples are important to build new risk models and to sustain existing ones. The choice of which populations to sample from, and how to conduct that sampling (i.e. desk or field) is directly related to the purpose of the models.

**Desk audits**
In the table below, we show some facts about the current process for desk audits and for the random samples for those processes.

| Tax type | Popu-lation size | Number of returns | Number of manual audits | Number of samples[5] | % sample from audit capacity | % sample from population |
|---|---|---|---|---|---|---|
| **Income tax (private persons)** | 10.4m | 10.4m | 490k | 23k | 4,7% | 0,22% |
| **Income tax (companies)** | 1.0m | 1.0m | 94k | 4,5k | 4,8% | 0,45% |
| **VAT negative** | 1.5m | 2.0m | 33k | 0,6k | 5,5% | 0,12% |
| **VAT positive** | 1.5m | 4.0m | - | 1,2k | | |
| **VAT nil** | 1.5m | 1.5m | 0 | 0 | - | - |
| **Company tax** | 1.5m | 650k | 50k | 0 | - | - |

The **income tax** (private persons) process has shown that a sample of approximately 5% of audit capacity is a reasonable balance between the various business needs.

The **VAT** sample is heavily weighted towards VAT positive as this is a brand new area for desk audits and modelling. In addition, it should be noted that here negative and positive refer to the balance of the return where a negative return means that a refund of VAT is being claimed. VAT desk audits look for receipts from the company that substantiate the negative (purchases) part of the return, so even for the VAT positive returns it is just the purchases part that is examined. Defining a way of desk auditing the positive (sales) part of the return is still to be done.

[5] The samples for income tax (companies) and VAT positive and negative are relatively new, having been started in January 2016.

For **company tax** desk audit capacity is around 50k and 5% of this 2.500. Additionally this is again the same population as VAT, which has a random sample of 1.800 so a desk audit random sample of around 2.000 seems sensible.

**Field Audits**

Most of the field audits are triggered by suspicious activity. In addition, 3.600 companies are included in the 'Random sample Companies'. These are 'whole company' or 'integral' field audits, where an experienced auditor examines some or all of a company's accounts and tax returns. This audit can encompass multiple years and multiple taxes.

| Tax type | Population size | Number of manual audits | Number of samples | % sample from audit capacity | % sample from population |
|---|---|---|---|---|---|
| **Cross-tax** | 1.5m | 35.000 | 3.600 | 10% | 0,24% |

There are three ways this random sample data could be used:

a) the results from this field random sample can be used for a future model that assesses the company as a whole and is not used to trigger examination of particular returns or individual taxes. In order to integrate the results of the audit with other data about the company the result would need to be tied to specific tax year(s).
b) If we wish to use the results of the 'Random samples Companies' for tax-specific models that are not triggered by a particular return, then we need the results from the audits to be recorded against those specific taxes[6] as well as specific year(s).
c) and if we wish to use this data as a random sample for a model that is used to select returns for desk audit the results must additionally be tied to specific returns.

The field audit results will contain information that is uncovered only by a thorough examination of the accounts and so a model built only on this data may well flag a company where a desk audit of a return would not show any risk. Since it is expensive to investigate every signal with a field audit signals of this type risk being ignored at the desk audit. Alternatively, since no risk model has 100% accuracy conducting a field audit for every signal risks wasting capacity.

[6] In addition the auditor would need to record which taxes they had examined, as a null result for a tax that has been checked is usable data, whilst a null result for a tax when we don't know if it was checked or not is not reliable.

Additionally, the field audit investigates many months, sometimes years, of a company's data, and usually does so over multiple taxes, for instance company tax and VAT. Which means that the result of an audit of this type (whether the company has over- or under-paid tax) cannot be tied to a specific return, or even to a specific tax. So, for example, a VAT risk model built on this data would run the risk of repeatedly flagging VAT returns from a company, or of flagging returns where a desk audit of *that particular return* or *that particular tax* would show no risk as the risk is spread over many returns or another tax or in all taxes taken together.

So while this 'full company' data can be used to supplement the data used to build the current risk models for companies, it is not sufficient in itself.

In general: for a sample to be useful for a risk model, it must sit within the business process that the model is used in. So a risk model that is used to flag VAT returns needs a sample, ideally a random sample, that is based on investigating VAT returns.