Digital forensic sampling

The application of statistical sampling in digital forensics

Authors: Robert-Jan Mora and Bas Kloet
Company: Hoffmann Investigations, Almere The Netherlands
URL: http://en.hoffmannbv.nl
Date: 27th March 2010
Version:1.0

Table of contents

1 Introduction

In this paper we would like to address a few problems that we encounter in the digital forensic field, in general, which probably will get worse if our methods do not get smarter soon. A few problems that the digital forensic community has to deal with are:

- The amount of data that needs to be investigated in cases increases every year;
- Forensic software is unstable when processing large quantities of data;
- Law Enforcement has a huge backlog in processing cases in time;
- More and more pressure is placed on digital forensic investigators to produce reliable results in a small amount of time.

So what can we do to be more effective and investigate the right data at the right time? In this paper we would like to propose a solution based on the technique of random sampling, which can be applied to the working field of digital forensics. The goal of this paper is to explain:

- when and why random sampling might be useful in a digital forensic investigation;
- present the reader with background information on relatively straightforward random sampling techniques;
- describe a number of cases where random sampling might be used to drastically reduce the amount of work required in a digital forensic investigation, without a significant (negative) impact on the reliability of the investigation.

To the authors' knowledge, the application of random sampling in digital forensics is very minimal[1], even though random sampling is often used in other forensic fields. For example the use of statistics to accurately estimate the quantity and quality of illegal drugs: if a large amount of amphetamine tablets is found, then how can we determine if every tablet contains amphetamine, without testing every tablet? By using statistical sampling techniques to select an appropriate sample for testing, we can make a reliable estimate of the quantity and quality of the total population of tablets. Also, statistical sampling techniques are used in financial auditing or in fraud investigations to detect fraud in large populations.

One of the reasons for writing this paper was a news article in the Dutch media about a child pornography case. A suspect had 49.500 pictures with child pornography on his computer.[2] There are a lot of different techniques to detect known child pornography, for example by known hash-sets or skin-tone detection techniques, but a lot of unknown material still has to be reviewed by a certified investigator, to see if it fits the criteria for child pornography.

After reading the article we had the following questions:

- With what kind of certainty had been determined that exactly 49.500 child pornographic pictures were found on the computer of the suspect?
- How long did the investigation took before this could be determined and at what cost?
- Does the exact amount of child pornographic material found on a suspect's hard drives directly influence the length of their sentence?

Since the number of files on hard drives are increasing every year, a smarter method of investigating these populations has to be established. This paper presents a novel solution which is based on random sampling methods applied in digital forensic investigations. The paper is organized as follows:

- Section 2 covers sampling basics, sampling size and sampling techniques;
- Section 3 addresses the application of sampling in different forensic investigations;
- Section 4 shows an algorithm of a simple random sampling application based on file extensions;
- Section 5 contains the conclusion.

2 Sampling basics

In digital forensics we collect data from suspects and analyze the data acquired from their computers or cellular devices. In a lot of investigations we need to examine large volumes of data of a specific kind. In an ordinary child pornography investigation the examiner has to review tens of thousands of pictures or movies and assess if they fit the criteria of child pornography.

If we take the selection from the investigation in the introduction of this paper, then we have 49.500 pictures which contain child pornography. Let's assume that the total number of pictures was 100.000. In statistics the total collection of elements of which we can take a sample is called *population.* One picture in the example above is called a *case of element.* If the examiner reviewed every case of element (thus 100.000 pictures) this is called a *census.* But if we choose to select some of the cases of elements using a specific method this is called a *sample.* There are many sampling techniques. In this paper we will only discuss the random sampling technique.

2.1 The necessity for sampling

In the introduction of the paper we've given some arguments why, with the ever growing population

[1] S. Garfinkel, A.J. Nelson Fast disk analysis with random sampling, www.simson.net 2010
[2] http://www.meldpunt-kinderporno.nl/news/?v=2&lid=1&id=641&cid=1. Website checked 8th March 2010

of files and the demand for quicker results, it's not quite feasible to do a complete review of all material in some digital forensic investigations. Sampling is a good alternative for a complete census if [3]:

- researching of the entire population is not possible in practice;
- budget limitation makes it impossible to examine the total population;
- time limits makes it impossible to research the entire population;
- all data has been collected, but you need to produce results quickly.

With sampling you can reliably use observations about the sample to make a statement about the entire population. We believe that the problems that we currently have to deal with in a lot of digital forensic investigations, are a good reason to look at the possibilities of using sampling in these cases.

2.2 Determine sample size

Before we can take a sample in a digital forensic investigation, we need to determine the sample size. The sample size has to do with a number of factors, including the purpose of the study, population size, the risk of selecting a bad sample and the allowable sampling error. The examples and definitions in this section are based on a paper about determining the sample size. [4]

2.2.1 The level of precision

The level of precision is sometimes called the sampling error. This is the range in which the true value of the population is estimated to be. This value is usually expressed in percentages (+/- 5%) that need to be determined by the investigator before sampling. Because the level of precision can have a significant effect on the sample size with a certain confidence level.

So if an digital forensic examiner finds that 85% of the JPEG files in the sample are classified as pornography, and determined the level of precision at 5%, then the examiner can conclude that between 80% and 90% with a certain confidence level of the entire population of JPEG files will almost certainly contain pornography.

2.2.2 The confidence level

“The confidence level or risk level is based on the ideas encompassed under the Central Limit Theorem. The key idea encompassed in the the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value.”[5]

This means that if 95% is the selected confidence level 95 out of the 100 samples will have the true population value within the range of precision specified earlier. In practice a 95% confidence level with a +/- 5% precision rate is assumed reliable.

2.2.3 Degree of variability

The degree of variability in the attributes being measured, refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to obtain the given level of precision. The less variable a population, the smaller the sample size. The level of variability is expressed using the 'proportion' or 'P'. A proportion of 0.5 (or 50%) indicates the greatest level of variability, more than either 0.2 or 0.8. This is because 0.2 or 0.8 indicate that a large majority do not or do, respectively, have the attribute of interest. Because a proportion of 0.5 indicates the maximum variability in a population it is often used in determining a more

---

[3] M. Saunder et al, Research methods for business students 03 edition, 2004
[4] Glenn D. Israel, Determining Sample Size, document PEOD6,1992.
[5] Ibidem

conservative sample size, that is, the sample size may be larger than if the true variability of the population attribute were used. In this paper we will use formulas that assume a proportion of 0.5, so we can ignore the level of variability without choosing overly optimistic sample sizes.

2.3 Using formulas to calculate a sample size

There are several methods for determining the sample size. In this paper we will present a simple formula from Yamane to determine the sample size.[6] This formula can be used to determine the minimal sample size for a given population size.

The formula from Yamane is:

$$n = \frac{N}{1 + N(e)^2}$$

Where:

n = sample size
N = population size
e = the level of precision

This formula assumes a degree of variability (i.e. proportion) of 0.5 and a confidence level of 95%.

Example 1 shows an example where the population of some sort is 2000 and where a 5% level of precision is required. The sample to be examined by the investigator is 333 items or objects.

$$n = \frac{N}{1 + N(e)^2} = \frac{2000}{1 + 2000 \times (.05)^2} = 333$$

Example 1.

If we want a confidence level of 0.95, then the statistical tables tell us z is 1.96. If we substitute this in the above formula, we get:

$n = \frac{0.96\mathrm{N}}{0.96 + N(e)^2}$, which turns into the Yamane formula if we round 0.96 to 1.

In section 3 of this paper the Yamane formula will be applied to several digital forensic investigation scenario's.

Using the Yamane formula, we can easily determine the minimal sample size that we have to investigate for any given population size. The downside to this formula however, is that it gives us at most a confidence level of 95%. If we want a higher (or lower) confidence level than 95%, then we will have to use the original version of the Yamane formula. To get this formula, we start with the original formula that the above Yamane formula is based on:[7]

---

[6] Yamane, Taro. 1967. Statistics, An Introductory Analysis, 2nd Ed, New York: Harper and Row.

[7] Yamane, Taro. 1967. Statistics, An Introductory Analysis, 2nd Ed, New York: Harper and Row, page 258.

$$n = \frac{z^2 P(1-P) N}{z^2 P(1-P) + N(e)^2}$$

Where:

n = sample size
N = population size
e = the level of precision
z = the value of the standard normal variable given the chosen confidence level (e.g. z = 1,96 with a CL =95 % and z= 2,57 with a CL = 99%
P = the proportion or degree of variability

If we assume a proportion P of 0.5, then this formula can be simplified to:

$$n = \frac{0.25(Z^2) N}{0.25(Z^2) + N(e)^2}$$

So what do we do when we want a 99% confidence level with a population of 2000? By using the statistical tables we can determine that we then have a z-value of 2.57, which gives us the following outcome:

$$n = \frac{(2.57)^2 \times 0.25 \times 2000}{(2.57)^2 \times 0.25 + 2000 \times (0.05)^2} \approx 496$$

The above formulas are all based on the notion that we want to perform some investigation on a sample, and use the results to say something about the entire population. But what if we have a large dataset in which we have a relatively small number of items that are of interest, for example to determine if fraudulent transactions are present in a population. If we have 20.000 transactions and we assume that from those transactions only 100 transactions are fraudulent, then we can be reasonably certain that the likelihood of detecting the fraudulent transactions with a sample size of 50 is not large. The following formula can be used to determine exactly *how* likely it is that we detect a fraudulent transaction: [8]

$$P = 1 - (1 - n/N)^E$$

Where:

n = sample size
N = population size
P = probability of selecting an error in the sample
E = number of items, e.g. a fraudulent transaction

[8] David Coderre, Computer-aided fraud, prevention & Detection, pag 224-225

So if we assume that in a population of 20.000 transactions only 100 are fraudulent and a sample of 50 is selected, the probability of finding a fraudulent transaction in the selected sample is:

$$P = 1 - (1 - 50/20000)^{100} = 22\%$$

But if we increase the sample size to 400 the probability of finding at least one fraudulent transaction is 87%.

$$P = 1 - (1 - 400/20000)^{100} = 87$$

2.4 Why random sampling

After the sample size has been determined by using the appropriate method, the preferred method is that that samples are randomly selected. In this way every element in the population has the same chance of being selected. Therefore sampling is done “without replacement”. This method avoids choosing a case of element more than once. The randomly selected sample is then called representative for the entire population.

3 Application of random sampling in different cases

In section 2 the basics of statistical sample were covered. In this section we will apply the statistical sampling method in several types of investigations that we regularly encounter in digital forensics. The sampling method is especially useful in investigations where a lot of data has to be reviewed. In this paper we will apply the method on pornographic cases and to detect fraudulent communication on hard drives. Also two examples of random sampling as a triage method and to reduce data in investigations, are explained.

3.1 Pornographic cases

One of the main fields of digital investigations within the Law Enforcement community is the investigation of child pornography. As a consequense, there are backlogs in dealing with those cases within the LE community. While a lot of efforts are done using known hashsets or using skin-tone detection techniques, a lot of the pornographic content has to be inspected manually.

So if we would take the example of the case referenced earlier in this paper with 49.500 pictures of child pornography found on a hard drive of a suspect, it would normally consume a lot of time and effort for a digital forensic investigator to review all the material including non pornographic material, the *census.* Our personal experience has revealed that reviewing such large quantities of data reviewing that is prone to human error. Furthermore if large quantities of such data are found on a suspect's hard drives this is not affect the sentence of a court of law in the Netherlands. This is also an argument to keep the material to be reviewed to a minimum since reviewing so many images could have a negative psychological impact on the forensic investigator.

So let's determine the sample size on a hard drive with 60.000 JPG pictures on it. These quantities of JPG files represent realistic numbers found in regular cases. Before we proceed we will need to determine the precision and confidence level. We will present the case with two confidence levels, of 95% and 99%, with a level of precision of 5% in both cases.

$$n = \frac{N}{1 + N(e)^2} = \frac{60000}{1 + 60000 \times (.05)^2} = 397$$

Example 2. Sample size of JPG files with a confidence level of 95% and a level of precision of 5%

With the confidence levels used in example 2, only 397 JPG files have to be reviewed by an investigator. Once the sample has been reviewed by the investigator and he or she finds that 70% in the sample contains child pornography he or she may conclude with a confidence level of 95% that between 65% and 75% of the JPG files contain child pornographic content.

$$n = \frac{1.65\mathrm{N}}{1.65 + N(e)^2} = \frac{1.65 \times 60000}{1.65 + 60000 \times (.05)^2} = \frac{99000}{151.65} = 653$$

Example 3. Sample size JPG files with a confidence level of 99% and a level of precision of 5%

So once the confidence level is rated higher the sample size increases. Still only 1.1% of the entire population has to be reviewed.

What if another suspect has a special folder for storing his pornographic JPG files? A count of these pictures shows that he has 100.000 JPG files stored in this special folder. How big would the sample size need to be with a 95% confidence level and a level of precision of 5% that have to be reviewed after they have been randomly selected? The answer is that only 398 pictures would have to be reviewed by a forensic investigator to conclude how the entire population looks like between the precision percentages. Wouldn't this be a nice feature in the gallery view of certain forensic applications, like EnCase and Forensic Toolkit?

Now let's assume that you are planning the search and seizure of a suspect's computers in a child pornography case. In this case you have very strong indications that the suspect is a high ranking member of a paedophile ring. Only the suspect lives in a house with several other persons who also have personal computers. Is it possible to make a well founded decision on site on which computer you'll need to acquire? Usually the police in the Netherlands decided to take all the computers with them, which they will then have to process at the office.

We could try to use random sampling as a triage method. In this case we assume that within the population of the 60.000 JPG files at least 500 pictures with child pornographic content are present. So if we take a sample size of 300 pictures that are randomly selected we have a probability of 92% of finding at least one picture with child pornographic content.

$$P = 1 - (1 - 300/60000)^{500} = 92\%$$

Example 4. Probability of detecting at least one JPG file with child pornographic content

The larger the sample size, for example a few hundred more, the higher the probability of detecting at least one JPG file on the other computers. If you do not detect them on site by using this triage technique, than you can take the risk of leaving the other personal computers on site, which in turn will not have to be investigated during the investigation. This could reduce backlogs significantly.

3.2 Detect fraudulent correspondence

Within e-discovery, compliance or fraud cases usually a lot of data needs to be reviewed. A lot of information is already gathered before the investigation is launched. Usually the data that gets processed are office documents containing correspondence and e-mail messages coming from network storage locations or hard drives. In ordinary investigations the processing of data takes up a lot of time, from several days to sometimes a week. Currently the forensic applications do not seem to handle the amount of data that needs to be processed within these ordinary cases. This problem will only get worse since the data quantities raise every year. In our experience in the investigations where the data was entirely reviewed by examiners based on the initial information, a lot of relevant files were found, normally 40 to 100 files.

So if we have a population of 10.000 Word files and we randomly select 600 Word files (case elements) we have a 92% probability that we detect at least one fraudulent Word file. The 600 files can be easily be processed and reviewed based on the initial information on site. See example 5.

$$P = 1 - (1 - 600/10000)^{40} = 92$$

Example 5. Probability of detecting at least one fraudulent Word file

4 Example program

This algorithm describes a program which takes the paths of a set of JPEG files as an input, and returns a subset of this list which contains a random sample of the paths in the original list. The size of the returned sample is based on the Yamane formula described in section 2.3 with the confidence level of 95%. For this example a precision rate of 5% percent is chosen, but this percentage can of course be easily adjusted in the algorithm.

```
# 'jpeg_list' is a variable that contains the list of JPEG paths. It is
# assumed to be available at the start of the algorithm.

# Set the precision rate level. Instead of being hardcoded, this can of course also
# be provided as a parameter to the algorithm.
precision_rate = 0.05

# Determine the population size
population_size = length(jpeg_list)

# Calculate the sample size using the Yamane formula with a 95% confidence level
sample_size =
  population_size / ( 1 + (population_size * precision_rate) ** 2)

# Take 'sample_size' random elements from the 'jpeg_list' and puts them in the
# 'sample' list
sample = []
i = 0
while i < sample_size
  rnd = random number between 1 and (population_size - i)
  remove the element at position 'rnd' from 'jpeg_list' and put it in 'sample' list
  i = i + 1
end

# 'sample' now contains a randomized sample from the 'jpeg_list'
```

The above program can of course be used to take a sample of any list of files.

5 Conclusion

In this paper we used a simple method for determining relevant sample sizes for different kinds of investigations and introduced a simple random sampling technique to obtain a sample from a certain population.[9] Although statistical sampling is used in fraud detection, drug sampling and voting, the use of statistical sampling is absent within the main forensic investigation applications like EnCase and FTK.

By using these statistical sampling techniques, results can be produced more quickly and with a reduced chance of human error than by simply reviewing all of the material, while still maintaining a known degree of confidence and precision. Statistical sampling provides for more credible support for conclusions, which can gain time and can help to reduce backlogs in the digital forensic field.

[9] There are more complicated sample formulas wich can be used to reduce the sample size, for example by using Bayesian methods.

Investigating a sample and basing conclusions on that sample involves taking a certain amount of risk, but it's our belief that the same risk is taken anyway by reviewing all the material in certain digital forensic investigations

5.1 Acknowledgements

We would like to thank our collegue and friend Marcel Westerhoud for his expertise and his critical review of this paper.