Triaging Evidence

Statistical Sampling in DFIR

Ray Strubinger
DFIR Managing Consultant

VERSPRITE

Your Speaker – Ray Strubinger

- Managing Consultant, Digital Forensics & Incident Response at VerSprite
  - Background in IT & Information Security Operations
  - Industry experience in financial services, government, higher education, healthcare, software development and consulting
  - Certifications in forensics, auditing and incident management
  - Led or participated in over 100 cases
    - Hacking, fraud & assorted white collar crimes
    - Large & small organizations

Goal

Understand how sampling may be used to triage files and identify areas that merit more in-depth analysis.

Why sample?

- It's all about…

Challenge – Information Overload

Challenge - Resources

- Limited time, people & other resources
  - Imagine 10,000 files
  - 1 second to review each file – about 3 hours
  - 10 seconds per file – about 30 hours
  - 20 seconds per file – about 60 hours
  - 30 seconds per file – about 80 hours

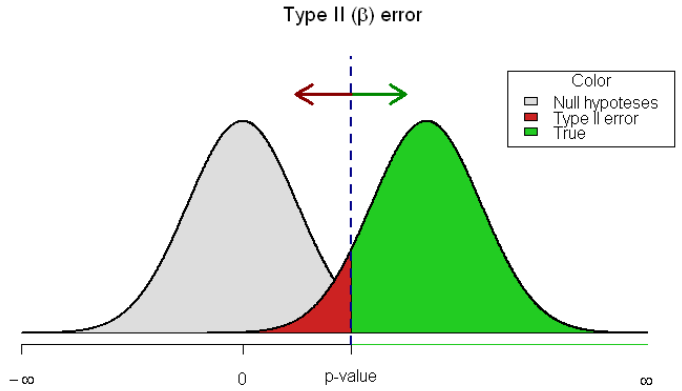
Addressing the Challenges

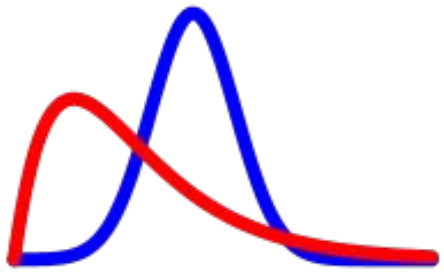
- Brute force – add people until the analysis time falls to an acceptable level
  - Does the case effectively scale with people?
  - Are people available?
- Do nothing – a potentially legitimate choice depending on the circumstances.
  - Output/reward verses effort may not be deemed to be worth it

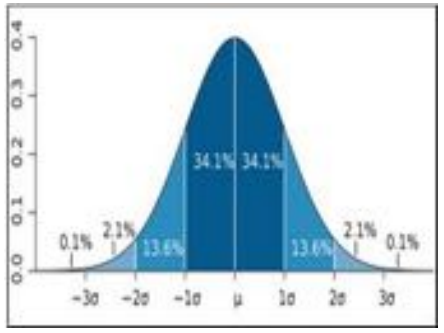
Addressing the Challenges

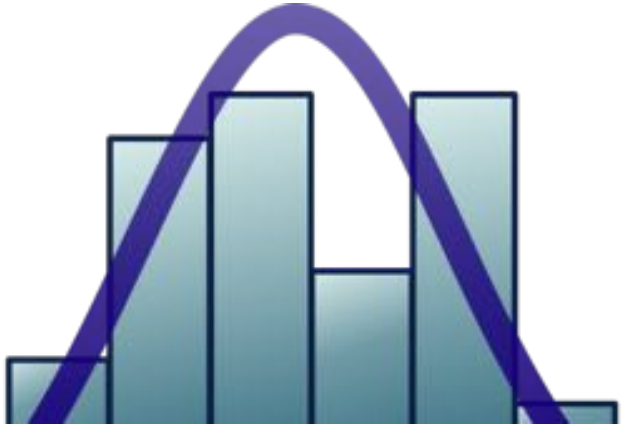
- Leverage technology – store & index files
  - Does the file type index?
    - Can we leverage key words or frequency analysis?
  - How much data do we have to index?
  - How long will it take to index the data?
- Improve the Signal to Noise ratio – use a technique to guide our focus and efforts
  - Ideal technique provides objectivity, rigor & repeatability

Statistics & Sampling









How does sampling work?

- Central Limit Theorem (greatly paraphrased)
  - Large random sample of files is representative of the remaining files
    - If 10% of the files in the sample fit the assessment criteria we can infer that 10% of the files in the larger collection will too
  - Sample size is calculated to provide a certain level of confidence for a given margin of error
  - References are included if you want a deeper dive on the theory

What's the approach?

- When the number of files to review is finite
  - Cochran-Yamane formula may be used

$$n = N/(1 + Ne^2)$$

  - n – number of files (samples) to review
  - N – number of files available for review (the population)
  - e – margin of error, usually 0.05
  - A 95% confidence level with a 5% margin of error is common

What does the formula tell us?

- Table shows 500 files are more than enough to be a “large” sample
- N – file population
- n – number of files to sample
- Why 500 files?
  - Need to review at least as many files as shown in column “n”
  - 500 provides a cushion

N	n
1000	286
2000	333
4000	364
5000	370
10000	385
20000	392
40000	396
50000	397
100000	398
1000000	400
10000000	400

How does the sample impact time?

- Random sample of 500 files
  - 10 seconds per file – about 1.5 hours
  - 20 seconds per file – about 3 hours
  - 30 seconds per file – about 4 hours
  - 60 seconds per file – about 8 hours

How do we use the technique?

- Screen by file type
- Generate a list of files
- De-duplicate via hash value
- Separate known & unknown
  - Vendor provided files may not be interesting
  - Use a “known good” list as a filter

How do we use the technique?

- Context of the case may suggest initial file types
  - Financial issues may favor documents & images
  - Administrative reviews may favor images over documents
  - Threat hunting or hacking may favor binaries over other file types
- Files types may include:
  - Images (JPGs, PNGs, GIFs)
  - MS Office docs (Word, Excel, PowerPoint, Access)
  - PDFs
  - Email
  - Archive formats (zip, rar, bzip)

Using the technique

- Now that we have a file list of all files…
- Randomize the file list
  - GNU sort or MS Excel
- Select 500 files for review
- Collect the 500 sample files
  - Access or copy the files so they may be reviewed
  - Consider transferring the files to their own area for export to other apps or external groups

Using the technique

- Review the sample of files
  - Identify files that fit the criteria of the case
- Use the results
  - If 50 of the 500 files fit the criteria of the case – 10% of all the files in the larger file collection will fit the case's criteria within 5%
  - By checking 500 files we are at least 95% confident that the larger file collection is within 5% of what the sample indicated
  - Take action based on the results
    - Review the item in greater detail
    - Reject the item and move to another
    - Adjust approach

Summary of Numbers, 1 - 10,000		
Numbers Ending	Count	Percentage
Single Digit	1000	10
Two of a Kind	100	1
Three of a Kind	10	0.1
Four of a Kind	1	0.01

Demo with numbers

- Try this at home!
- Generate a list of numbers, 1 – 10,000
  - Collection mimics a population of files
  - Collection also functions as a standard
- Script listed in References
  - Generates the number file
  - Creates 1000 files with 500 random numbers randomly selected from the number file

Demo with numbers

Test of 1000 Files		
Numbers Ending in	Count	Percentage
7	1000	100
77	996	99.6
777	383	38.3
7777	48	4.8

- Simulation of 1000 cases
- Select 500 numbers at random from the list of 10,000 numbers
- Notice the row that lists “77”
  - Only 1% of the 10k numbers end in 77
  - Any two digit number may be selected and still have a 1% rate of occurrence. 33, 44, or 99 work as well as 77.

Test of 24 Files		
Numbers Ending in	Count	Percentage
7	24	100
77	24	100
777	4	17
7777	1	4

Demo with numbers

- Simulation of 24 cases
- Select 500 numbers at random from the generated list
- Notice the “77” row again – all 24 sample files have at least one number ending in 77
  - 17 out of 24 sample files – 71% had at least 4 numbers ending in “77”
  - 4 out of 500 is approximately 1%

Test of 1 File			
Numbers Ending In	Line Count	Sample Percentage	Population Percentage
7	55	11	10
77	6	1.2	1
777	0	0	0.1
7777	0	0	0.01

Demo with numbers

- Simulation of a single case
- 500 numbers selected at random
- Here we compare the sample to the actual population percentages
- What do we make of this?
  - Even when the occurrence rate of "interesting" files is as low as 1% the odds of detection are in our favor

Thank you

Ray Strubinger
rays@versprite.com

Slides & Script: www.versprite.com/SANSDFIR

References

- Central Limit Theorem - http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability12.html
- Cochran-Yamane - https://www.tarleton.edu/academicassessment/documents/Samplesize.pdf