If Error Rate is Such a Simple Concept, Why Don't I have One for My Forensic Tool Yet

By

James Lyle

Presented At

The Digital Forensic Research Conference

DFRWS 2010 USA Portland, OR (Aug 2nd - 4th)

If Error Rate is Such a Simple Concept, Why Don't I Have One for my Forensic Tool Yet?

Jim Lyle

National Institute of Standards and Technology

Disclaimer

Certain trade names and company products are mentioned in the text or identified. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Introduction

- Daubert – criteria to assess admissibility of scientific testimony
  - Tested
  - Peer review
  - Error rate & controls
  - General acceptance
- The first idea (using tool test results) for establishing an error rate doesn't work.

First try for an Error Rate Fails

- Consider disk imaging . . .
  - Let n be total bits acquired
  - Let k be number of incorrectly acquired bits
  - Then k/n looks like an error rate.
- But, how to determine n & k is hard.
- Doing lots of acquires may not get a representative sample of drives that might be imaged.

Outline

- Typical errors seen during testing
- Measurement & Statistical Errors
- Sources of Errors
- An Example
- Establishing Error Rates
- Summary

Disk Imaging Behaviors

- Some sectors omitted
  - 1024 sectors for Quantum Sirocco (SafeBack)
  - 5040 sectors for Quantum Sirocco (EnCase 3)
  - 1 sector if drive has an odd number of sectors (dd Linux)
  - Last 8 sectors of NTFS logical drive (FTK)
  - Last sector of NTFS logical drive (EnCase 4, 5 & 6) and seven sectors prior to last sector are a repeat from earlier in the image.
  - Sectors around a faulty sectors replaced by zeros
  - HPA & DCO

Testing a Hypothesis – Does entity X have attribute A?

- Statistical process
- A Matrix of possibilities

Test Result	Reality	
	X has A	X does not have A
X has A	Accept	False Positive aka Type I Error
X does not have A	False Negative aka Type II Error	Reject

Error rate for each type of error is the probability of the error occurring.

Sources of Error

- The theory of measurement error identifies two classes of errors: measurement (random process) & systematic (non-random)
- For forensic tools that implement some algorithm . . .

1. An algorithm may have a theoretical (random process) error rate
2. An implementation of an algorithm may have systematic (non-random) errors
3. The execution of a procedure may have a blunder that affects the result

- Daubert is mostly interested in the first two.

Error Source Example

- Hashes or checksums (with useful attributes) can be computed for a file.
  - Same files have the same hash
  - A different hash means files are different
  - However, the same hash is possible for different files
- Hashes or checksums can be used to determine if:
  - A file has changed, or
  - If two files might be the same with some error rate.

An Algorithm To Compare A Pair Of Files With Only One File

- A hash or checksum can be used to determine if any file in a set of files match a given file.

1. Let c be the hash of the given file
2. For each file, f, in the set …
    1. Compute, h, the hash of f
    2. Compare c to h
    3. If c matches h, then declare c equals h

- Hashes can collide (two different files with same hash)
- The error rate (type I) of file matches is related to the size of the hash (number of bits)
- The error rate (type II) for identifying two identical files as different is zero.

Comparing Randomly Selected Files

Chance of hash or checksum for a random file matching a given value

Algorithm	Chance of Collision
CRC-16	1 in 32,768
CRC-32	1 in 2,147,483,648
MD5 (128 bits)	1 in 170141183460469231731687303715884105728
SHA-1	1 in 2^{159}
SHA-256	1 in 2^{255}

Implementation Errors

- A variety of implementation errors are possible, some are quite subtle.
  - One common error occurs as follows:
    - Hash algorithm is implemented in a UNIX environment. It works for any file.
    - Same program is moved to MS Windows environment. It works fine for any *binary* file, but computes a different (wrong) value for any *text* file (Windows adds a character to the end of each line of text).

What is the error rate?

- In the science of measurement error analysis an implementation error is called a *systematic error.*
- The distribution of text and binary files varies from computer to computer.
- There is no random distribution to the manifestation of the error.
- The implementation error is triggered only under some set of conditions.
- A tool may have implementation errors, but the algorithm being implemented has a statistical error rate.

Human Errors

- Human errors (blunders) occur
- Difficult to quantify
- Good processes have built in checks to detect blunders

Error Rate for Disk Imaging

- Forensic tools often have multiple requirements.
- Each requirement may generate a separate error rate.
- Separate the algorithm from the implementation.
- Algorithm is . . . Read and make a copy of every accessible sector on the drive. The error rate is zero.
- The implementation may have a many different systematic errors.
- Alternate algorithm . . . Add an attempt to read additional (not accessible) sectors – Unknown error rate.

Other Error Rates

- Write blocking
- String Searching
- File Recovery and Carving

Summary & Observations

- Distinguish between intended algorithm and actual implementation
- Algorithm may have an error rate (statistical in nature)
- Implementations have systematic errors
- Most digital forensic tool functions are simple collection, extraction or searching operations with a zero error rate for the algorithm.
- Tools tend to have minor problems, usually omitting data, sometimes duplicating existing data.
- An implementation's systematic errors can be revealed by tool testing.
- To satisfy the intent of Daubert, tools should have the types of failures and triggering conditions characterized.

Project Sponsors (aka Steering Committee)

- National Institute of Justice (Major funding)
- FBI (Additional funding)
- Department of Defense, DCCI (Equipment and support)
- Homeland Security (Major funding)
- State & Local agencies (Technical input)
- Internal Revenue, IRS (Technical input)
- NIST/OLES (Program management)

Contact Information

Jim Lyle

jlyle@nist.gov

http://www.cftt.nist.gov

Sue Ballou, Office of Law Enforcement Standards
Steering Committee representative for State/Local Law Enforcement
Susan.ballou@nist.gov