Figure 2: Binomial Distribution

Using Bayesian Statistical Methods to Determine the Level of Error in Large Spreadsheets

Leslie Bradley and Kevin McDaid
Software Technology Research Centre (SToRC)
Dundalk Institute of Technology, Ireland

Introduction

Spreadsheets, Spreadsheets. Spreadsheets

- ❖ Who – Auditors, Accountants, Managers
- ❖ What – Accounts, Budgets, Databases
- ❖ Why – Easy to use, Holds large amounts of data
- ❖ Where – Finance sector, Business, Science
- ❖ When – Everyday
- ❖ How – Created with little controls and guidelines
- ❖ Result – A high dependence on an application with few controls, and a lack of guidelines and best practices

Errors

- ➢ The Nevada city budget showed a deficit of $5 million dollars because the spreadsheet was not updated. January 2006
- ➢ A cell entry error cost Columbia Housing Authority $118,387, which was overpaid to landlords. February 2006
- ➢ Deliberate fraud, AIB losses of nearly $700 million dollars were hidden by a trader, John Rusnak. 2001 http://news.bbc.co.uk/1/hi/business/1805777.stm

[1]

BBC CATEGORIES TV RADIO COMMUNICATE WHERE I LIVE INDEX SEARCH Go

BBC NEWS

You are in: Business

Front Page
World
UK
UK Politics
Business
Market Data
Economy
Companies
E-Commerce
Your Money
Business Basics
Sci/Tech
Health
Education
Entertainment
Talking Point
In Depth
AudioVideo

Thursday, 7 February, 2002, 09:51 GMT

Rogue trader 'Mr Middle America'

It is unclear if John Rusnak has benefited financially

The words 'rogue trader' conjure up a piratical, swashbuckling image, of a hotshot here today and gone tomorrow.

But in the case of Allfirst employee and "Mr Middle America" John Rusnak, nothing, it seems, could be further from the truth.

According to his superiors at Allfirst, AIB's Baltimore-based subsidiary, he had been with the company since

"He seems like a really cool guy. His kids are cool. I'm shocked

AIB ROGUE TRADER

Irish bank hit by fraud

- Rogue trader meets FBI
- The fraud step-by-step
- How to lose $750m
- 'Mr Middle America'
- Stock market rocked
- Leeson's view
- Rogues' gallery
- Profile: AIB
- What are futures and options?

AUDIO VIDEO

- TV and Radio reports

See also:

- 22 Jun 99 | The Economy How Leeson broke the bank
- 22 Jun 99 | The Economy Leeson scandal 'could happen again'

Internet links:

- Allied Irish Banks
- Allfirst

BBC SPORT
BBC Weather

SERVICES
Daily E-mail
News Ticker
Mobiles/PDAs

City of London

An overview into the uses of spreadsheets in the city of London shows the level of involvement spreadsheets have in the finance sector, [2].

- Actuary spends 90% of their day in spreadsheet environment
- 256 column limiting financial modelling
- Spreadsheets greater than 1GB in size
- Unique formulas numbering from 1,000 – 10,000 and upwards

> "The time taken to review these models can range from twenty five hours to many hundreds, generating significant fee income for firms undertaking this work,"

Error Rates

Research has been conducted to measuring the level of error in spreadsheets. The cell error rate (CER) is the commonly used way of measuring the error. The CER is the percentage of cells in the spreadsheet that have errors, [3].

- 5.2% CER for a study involving 43 spreadsheets, [4, 5]
- Later study [6], used 50 spreadsheets
  - 1.79% CER for general errors
  - 0.87% CER for wrong results

Wrong result – an error that is an incorrect result
Poor practice – an error that gives the correct result

Issues

- Spreadsheets error is not universally defined
- Spreadsheets need to be fully audited to produce CER
- Do not incorporate external factors that influence errors

Research Question

Can a model be established to predict the cell error rate of large spreadsheets, based on expert knowledge and any available test data, to aid the decision on whether to test the spreadsheet?

❖ The CER is now defined as the probability of the cell containing an error.

Bayesian Model

- Combine expert knowledge and any available test data
- Expert knowledge or Prior information can be based on factors like:
  - Spreadsheet Complexity
  - Developer Skill
  - Company policy

Prior Information

- Prior information is represented as a beta distribution.
- Expert information provides likely error rate and an indication of the spread in the value
- From this parameters for the beta distribution (α, β) can be deduced.
  - α – number of prior error cells
  - β – number of prior error free cells

$$f(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_{\theta} \theta^{\alpha-1}(1-\theta)^{\beta-1}} \quad \text{where} \quad 0 \le \theta \le 1$$

Test Data

- Test data is represented as a binomial distribution.
  - x – number of error cells found during testing
  - n – number of cells tested

$$b(x;n,\theta) = \binom{n}{x} \theta^{x}(1-\theta)^{n-x}$$

$$\text{for} \quad x = 0, 1, 2, \ldots, n$$

Posterior Information

- Combination of prior information and test data produces posterior information which is represented as a beta distribution.
  - α + x – posterior error cell information
  - β + (n-x) – posterior error free cell information

$$f(\theta) = \frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+(n-x)-1}}{\int_{\theta} \theta^{\alpha+x-1}(1-\theta)^{\beta+(n-x)-1}} \quad \text{where} \quad 0 \le \theta \le 1$$

Decision Making

- Some Financial institutions may not conduct large spreadsheet audits if the level of error is below a certain value.
- Bayesian methods allow the calculation of a reliability which, returns the probability that the predicted CER stays below the acceptable CER (A-CER).
  - This provides a mathematical basis for the decision

Example

Suppose a company has a suite of large spreadsheets. The auditor wants to predict the level of error in the spreadsheets to determine if a full examination is required. One spreadsheet contains 303 unique formulas. After consultation with experts and discussion on the particular spreadsheet and developer, prior information was deduced. It indicates a defect error rate of 0.05 with standard deviation of 0.0217. The first 200 unique formulas are tested and results show 2 defect error cells. This test data is added to the prior to give posterior information which has mean 0.023 and standard deviation of 0.0087.

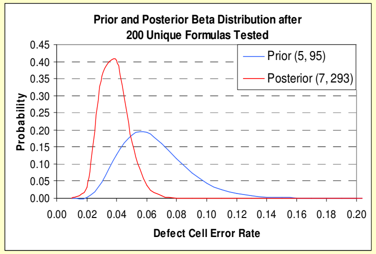

Figure 1: Prior and Posterior Distribution showing the likely Defect CER for the spreadsheet.

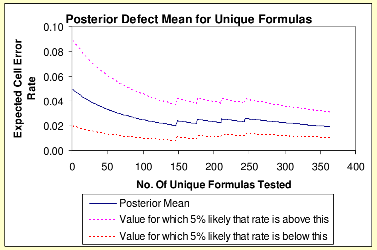

Figure 2: The expected Defect CER at each unique formula time stop.

Conclusions and Future Work

- ❖ Spreadsheet errors can have significant financial impact.
- ❖ Examination of large spreadsheet can be costly and time consuming
- ❖ Bayesian model incorporates expert knowledge and test data to predict CER.
- ❖ Provide mathematical basis for decision on whether a complete audit is required.
- ❖ Bayesian model is organisation-based to allow prior information to be used for other spreadsheet projects.

- ➢ Model to be evaluated using operational spreadsheets.
- ➢ Review the process to select the cells for the test data.
- ➢ Investigate the relationship between similar cells.

References

1. EuSpRiG, 10:55 a.m. February 25, http://www.eusprig.org/stories.htm
2. Croll, G.J. *The Importance and Criticality of Spreadsheets in the City of London*. in *Proceedings of the European Spreadsheets Risks Interest Group*. 2005. London, England.
3. Panko, R.R., and Halverson, Jr., R. P. *Spreadsheets on Trial: A Survey of Research on Spreadsheet Risks*. in *The 29th Annual Hawaii International Conference on System Sciences*. 1996. Hawaii.
4. Panko, R.R., *What we know about spreadsheet errors.* Journal of End User Computing Special issue on scaling Up End User Development, 1998. **10**(2): p. 15-21.
5. Panko, R.R., *What we know about Spreadsheet Errors Extended Version*, February 12, 2008, 2005, http://panko.shidler.hawaii.edu/SSR/index.htm.
6. Powell, S.G., Baker, K. R., and Lawson, B., *Errors in operational Spreadsheets*, March, 2008, 2007, http://mba.tuck.dartmouth.edu/spreadsheet/index.html,

Acknowledgements

This research is supported by the Institutes of Technology Ireland's Technological Sector Research: STRAND 1 Post-Graduate R&D Skills Programme.

For more information contact
Leslie Bradley
leslie.bradley@dkit.ie