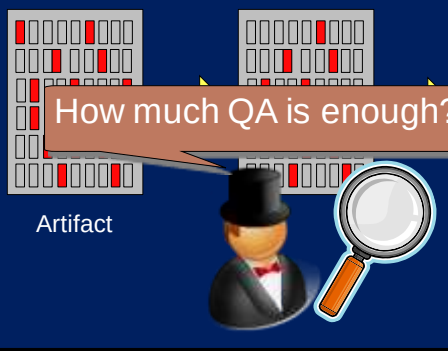
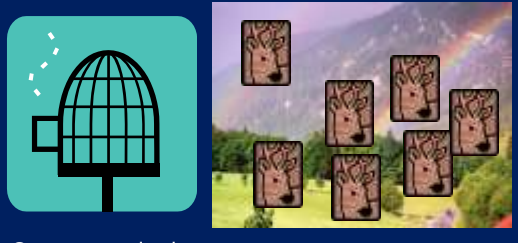
Discussion: "A Comprehensive Evaluation of Capture-Recapture Models for Estimating Software Defect Content"

by Lionel C. Briand, Khaled El Emam, Bernd G. Freimut, and Oliver Laitenberger

Bryan Robbins
CMSC 737, Fundamentals of Software Testing
November 5, 2009



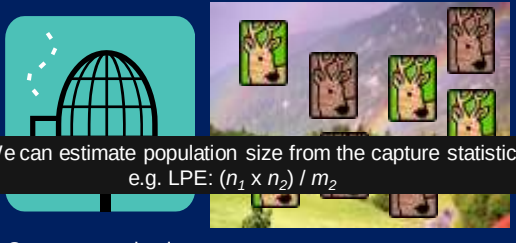
From Biological Sciences: Capture-Recapture



Capture n_1 animals

Population (Unknown Size)

From Biological Sciences: Capture-Recapture



Idea: We can estimate population size from the capture statistics, e.g. LPE: $(n_1 \times n_2) / m_2$

Capture n_2 animals, where m_2 are marked

Population (Unknown Size)

Capture-Recapture for Defect Estimation

- Application to QA Activities:
  - Use capture-recapture models to estimate total number of defects
  - Use total number of defects to inform QA decisions
- Many open issues:
  - Choice of C/R model?
    - Validity of C/R model assumptions?
  - Choice of estimator?

Briand, et al. 2000: Primary Contributions

- C/R Models tend to underestimate remaining defects
- Using a very small number of inspectors (< 4) leads to particularly inaccurate estimates
- Model calibration has a number of theoretical limitations
- The Jackknife estimator is recommended, and is based on a model that allows for different defect detection probabilities

Outline

- C/R Models
- Estimators
- Research Method
- Results and Analysis
- Issues
- Discussion: Validity for Software Testing

C/R Models

- Assumptions:
  - Only two trapping occasions
  - No animals enter or leave population between occasions
  - **All animals have an equal likelihood of being captured**

Observation: No model addresses the "interaction effect" – Inspector A is good at finding memory leaks, but poor at detecting race conditions.

Model	Detection Probability	Inspector Capability
M0	Same	Same
Mh	Different	Same
Mt	Same	Different
Mth	Different	Different

Estimators

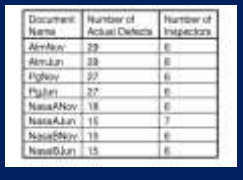
- Given the four C/R models
  - Need estimators based on sources of variation
  - Many estimators suggested in biology literature
  - Each requires different defect detection data
    - All data provided by a matrix of Defects x Inspectors

Model(Estimator)
M0(MLE)
Mt(MLE)
Mt(Ch)
Mh(JE)
Mh(Ch)
Mth(Ch)

Research Method

- Use an existing data set – Requirements inspection data from Basili, et al. 1996
- Create "virtual inspections" from data set
  - Vary number of inspectors and number of actual defects in document
- Compare model predictions to actual data for each virtual inspection
  - Relative Error (RE) for each model estimate
  - Describe central tendency and variability of RE
  - Report how often a model fails to produce estimate
- Select best model
  - Based on ordered hypotheses (!) using Dunn-Bonferroni tests

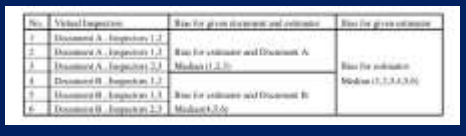
Varying Number of Inspectors and Defects

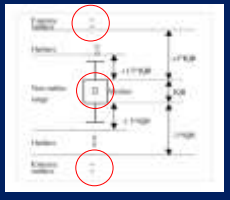


Document Name	Number of Actual Defects	Number of Inspectors
AllvNov	29	6
AllvJun	28	6
FlgNov	27	6
FlgJun	27	6
NasaANov	18	6
NasaAJun	16	7
NasaBNov	19	6
NasaBJun	15	6

- Virtual inspections created by choosing data of n inspectors from k actual inspectors
- Number of defects varied by sampling from all possible combinations of defects (hold number of inspectors constant)

RE Data from Virtual Inspections

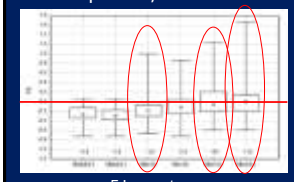




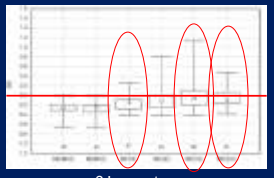
- Compute Median Relative Error as Bias (central tendency)
- Compute interquartile range (IQR) of RE (variability)
- Check for extreme outliers (variability)

Results: Varying Number of Inspectors

- Generally, models underestimate
- Ch estimators are most accurate, but most prone to extreme outliers
- Tendency for extreme outliers decreases as number of inspectors increases
- No estimator is reasonably accurate with less than four inspectors, but calibration may be able to help



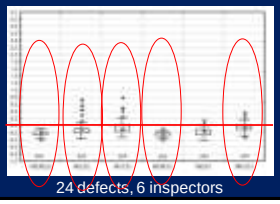
5 Inspectors



6 Inspectors

Results: Varying Number of Defects

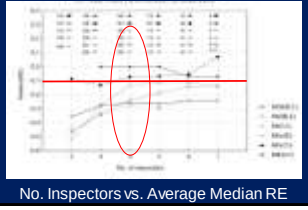
- Tendency for extreme outliers decreases as number of defects increases
- Median RE not greatly affected by number of defects
- Mh and Mth outperform M0 – Mt does not
- For Mh, Ch estimators have median RE closer than JE estimator



24 defects, 6 inspectors

Results: Selecting Thresholds

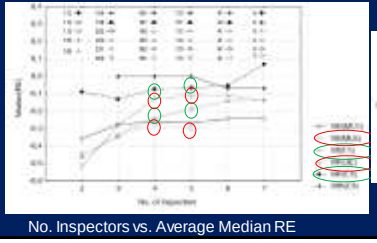
- Threshold for number of inspectors: 4 Inspectors
- Threshold for number of defects (see Figure 7):
  - Largest difference in median RE between 6 and 12 defects
  - After 12 defects, improvements are minimal.
  - For Mth, effect of number of defects minimal when using at least 6 inspectors



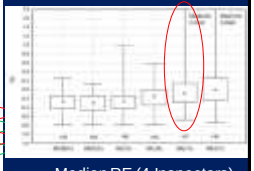
No. Inspectors vs. Average Median RE

Results: Best Estimators

- For Mt: Ch estimator outperforms MLE for 4 and 5 inspectors
- For Mh: Minimal difference for 4 or 5 inspectors, but for 4 inspectors, Ch prone to extreme overestimation



No. Inspectors vs. Average Median RE



Median RE (4 Inspectors)

Results: Most Appropriate Model

- Idea: Gathering data costs money, so adding data should significantly improve the model
- Compare estimates pairwise based on two ordered hypotheses
  - Significant Differences for 4 Inspectors: Mh vs. M0, Mth vs. Mt
  - For 5 Inspectors:
    - All comparisons significant except Mh vs. Mth
    - Mh(*) vs. M0 much more significant than Mt vs. M0
    - Mh(JE) vs. M0 much more significant than Mh(Ch) vs. M0
- Mh(JE) considered best model as measured by largest significant difference.

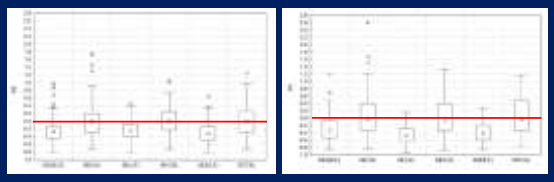
Results: Failure Rate

- Estimators rarely fail with at least 4 inspectors
- Mh(Ch) has highest failure rate across all conditions
- Mh(JE) has lowest failure rate across all conditions
- Provides more support for Mh(JE)

	All defects		4 Inspectors	
Estimator	< 4 inspectors	>= 4 inspectors	<12 defects	>= 12 defects
Mh(JE)	0.5%	0%	1.1%	0%
Mh(CK)	12.8%	3.9%	30.2%	14%
Mh(Ch)	6.2%	0%	15.7%	6.5%

Results: Calibrating Models

- Calibration improves median RE in all cases, but increases variability (particularly for 2 inspectors)



2 Inspectors 3 Inspectors

Issues

- Data set
  - Original experiment was evaluating PBR, which strives to minimize overlap (estimators depend on overlap)
  - Relatively small number of inspectors and defects
- Ordered hypotheses
  - All data fairly easy to obtain
  - Cost of simulation?
- Results
  - Mh(JK) still has very high variability, even for 5-6 inspectors
  - Walia, et al. 2008 found that 26 inspectors are required to stabilize the Mh(JK) variability (the worst among 12 models considered)

Discussion: C/R for Software Testing

- Data set was requirements inspections – what about defect estimation during testing?
- Related Work – Scott and Wohlin 2008 applied C/R to unit testing in a case study
  - Data matrix was Testers x Faults
  - Results from were "encouraging" (qualitative analysis)
- Can we use Test Suite x Defects ?
  - Randomly generated test suites of fixed size
  - What would Mt attempt to account for?