INVESTIGATION OF THE EFFECT OF THE NUMBER OF INSPECTORS ON THE SOFTWARE DEFECT ESTIMATES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Kaustubh Saxena

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

January 2012

Fargo, North Dakota

North Dakota State University
Graduate School

Title

INVESTIGATION OF THE EFFECT OF THE NUMBER OF INSPECTORS ON THE SOFTWARE DEFECT ESTIMATES

By
Kaustubh Saxena

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Gursimran S. Walia
Chair

Dr. Kendall E. Nygard

Dr. Charlene Wolf-Hall

Dr. Hyunsook Do

Approved:

January 30, 2012
Date

Dr. Brian M. Slator
Department Chair

ABSTRACT

Capture-recapture models help software managers by providing post-inspection defect estimate remaining in a software artifact to determine if a re-inspection in necessary. These estimates are calculated using the number of unique faults per inspector and the overlap of faults found by inspectors during an inspection cycle. A common belief is that the accuracy of the capture-recapture estimates improves with the inspection team size. This however, has not been empirically studied. This paper empirically investigates the effect of the number of inspectors on the estimates produced by capture-recapture models, by using inspection data with varying number and types of inspectors. The results show that the SC (Sample Coverage) estimators are best suited to software inspections and need least number of inspectors to achieve accurate and precise estimates. Our results also provide a detailed analysis of the number of inspectors necessary to obtain estimates within 5-20% of the actual defect count.

ACKNOWLEDGEMENTS

I would like to thank my major adviser, Dr. Gursimran S. Walia for his continued support, help and direction. I would also like to convey my gratitude to Dr. Kendall E. Nygard, Dr. Hyunsook Do, and Dr. Charlene Wolf-Hall for being on my graduate committee. I would also like to thank my family and friends who encouraged me to complete my paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Project managers and software developers manage the development process by monitoring the quality of the artifacts developed at each lifecycle stage. In the software engineering community, inspections are widely used to improve the quality of these artifacts, by enabling developers to detect faults early and avoid costly rework later [1]. In practice, however, the evidence suggests that the effectiveness of inspections varies widely [1, 2]. Furthermore, inspections only identify the presence of faults; they cannot certify the absence of faults or provide insight into how many remain post-inspection.

Project managers need objective information to help them decide when enough faults have been found that they can safely stop the inspection process. During a real project, a reliable estimate of the number of faults can aid mangers in determining the need for additional inspections. Among the various approaches available for estimating the number of faults (e.g., fault density, subjective assessment, historical trends, capture-recapture, and curve-fitting), capture-recapture is the most objective and appropriate method [2, 6].

Capture-recapture (CR) is a statistical method that was originally developed by biologists for estimating the size of wildlife populations. CR is used by repeatedly trapping (or capturing) a fixed number of animals, marking them, and releasing them back into the population. If the same animal is trapped during subsequent trapping occasions, it is said to have been recaptured. The size of the population is then estimated using: 1) the total number of unique animals captured across all trapping occasions, and 2) the number of animals that were re-captured. A higher percentage of recaptures indicates a smaller population [8, 14].

Using the same principle, the CR method can be used during the inspection

process to estimate the number of faults in an artifact. During an inspection, each inspector finds (or captures) some faults. If the same fault is found by more than one inspector it has been re-captured [2, 4]. The total number of faults is then estimated in a similar manner as in wildlife research, with the animals replaced by faults and the trappings replaced by inspectors. The difference between the estimated total number of faults and the faults already found provides an estimate of how many remain.

While biology and wildlife researchers have performed comprehensive evaluations of capture-recapture models using large data sets [16, 23], the studies in software engineering have been limited to relatively small data sets with a small number of inspectors and defects [2, 3, 10, 11, 13, 17, 18, 20, 24, 25]. In addition, often the capture-recapture studies in software engineering have used artifacts with seeded defects. Recommendations and comparisons of the software engineering findings with the biology and wildlife results are made based on those limited data sets. Since the capture-recapture models work based on the amount of overlap in the defects detected by different inspectors, it is unclear what effect a large number of inspectors will have on the performance of the capture-recapture models.

To calculate the estimates, the capture-recapture models use various mathematical estimators. Each estimator makes its own set of assumptions about the underlying data and therefore may produce different population estimates. In this paper, we evaluate the performance of different capture-recapture estimators for providing estimates with satisfactory accuracy and precision, on six different data sets that includes defect input data from different number of inspectors. Some of these estimators have not been studied in software engineering before, The number of inspectors used in these data sets varies from a minimum of six inspectors to a

maximum of seventy-three inspectors. Using this data set, the effects of team sizes ranging from two to seventy-three can be evaluated. We compare the results obtained from this study with the results and recommendations from previous software engineering research. We examine the effect of number of inspectors on the performance of different estimators. The results provide insight about the minimum number of inspectors required for achieving satisfactory estimates. Software developers and project managers can use these results to plan and manage inspections in their organizations.

Section 2 describes the basic principles of capture-recapture models and their application to software inspections. Section 3 discusses the background literature that motivated this study. Section 4 describes the design of the study used for evaluating the capture-recapture models. Section 5 describes the data analysis and results. Section 6 discusses the threats to validity. Section 7 discusses the relevance of the results and compares the results with previous results from software engineering and biology.

2. USE OF CAPTURE-RECAPTURE FOR DEFECT ESTIMATION IN SOFTWARE ENGINEERING

The use of the CR method in biology makes certain assumptions that do not always hold for software inspections. The assumptions made by CR method in biology include: 1) a closed population (i.e. no animal can enter or leave), 2) an equal capture probability (i.e. all animals have an equal chance of being captured), and 3) marks are not lost (i.e. an animal that has been captured can be identified) [15]. When using the CR in software inspections, the closed population assumption is met (i.e., all inspectors review the same artifact and it is not modified) and the assumption that marks are not lost is met (i.e. it can be determined if two people report the same fault). However, because some faults are easier to find than others and because inspectors have different abilities, the equal capture probability assumption is not met [2, 9].

To accommodate these different assumptions, four different CR models are built around the two sources of variation: Inspector Capability and Fault Detection Probability. Table 1 shows the four CR models along with their source(s) of variation. Each CR model in Table 1 has a set of estimators, which use different statistical approaches to produce the estimates. The estimators for each CR model used in this study are shown in Table 2. These estimators include estimators that have been evaluated in previous software inspection studies as well as new estimators from biology that have not previously been applied to software inspections (marked with an *).

Table 1 - Capture-Recapture Models

Model	Variation Source
$\mathbf{M_o}$	All inspectors have the same detection ability, and all defects are equally likely of being detected.
$\mathbf{M_t}$	Inspectors differ in their defect detection abilities, but all defects are equally likely of being found.
$\mathbf{M_h}$	Inspectors have the same detection ability, but defects differ in their probability of being found.
$\mathbf{M_{th}}$	Inspectors differ in their defect detection ability, and defects differ in their probability of being found.

Table 2 - Capture-Recapture Estimators

Models	Estimators
$\mathbf{M_o}$	Unconditional Maximum Likelihood Estimator (M_o-UMLE) [16]
	*Conditional Maximum Likelihood Estimator (M_o-CMLE) [8]
	*Estimating Equations (M_o-EE) [26]
$\mathbf{M_t}$	Unconditional Maximum Likelihood Estimator (M_t-UMLE) [16]
	*Conditional Maximum Likelihood Estimator (M_t-CMLE) [8]
	*Estimating Equations (M_t-EE) [26]
	Chaos Estimator (M_t-Ch) [5]
$\mathbf{M_h}$	Jackknife Estimator (M_h-JK) [4]
	*Sample Coverage (M_h-SC) [14]
	*Estimating Equations (M_h-EE) [26]
	Chaos Estimators (M_h-Ch) [6]
$\mathbf{M_{th}}$	*Sample Coverage (M_{th}-SC) [14]
	*Estimating Equations (M_{th}-EE) [26]

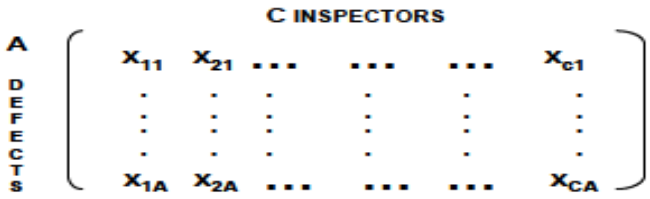

Figure 1 - Capture Recapture Data Input Matrix

The mathematical details of estimators are beyond the scope of this paper but can be found in provided references. The input data used by all the CR estimators is organized as a matrix with rows that represent faults and columns that represent inspectors as shown in Figure 1. A matrix entry is 1 if the fault is found by the inspector and 0 otherwise.

3. PREVIOUS EMPIRICAL STUDIES OF CAPTURE-RECAPTURE IN SOFTWARE ENGINEERING

Most CR research related to software inspections has focused on the basic theory and evaluation of CR models, with very little focus on the influencing factors involved [9]. The theory of CR for software inspections was introduced by Eick, et al. in an early study on the use of CR models for software inspections by applying them to real defect data from AT&T. They applied the maximum likelihood estimator for the M_t model to estimate the number of faults remaining in requirement and design artifacts. The estimates produced by CR were similar to the subjective opinion of the inspectors. A major result from this study was the recommendation (based on the inspection results) that an artifact should be re-inspected if more than 20% of the total faults remain undetected [4, 5]. This recommendation has been used by all subsequent CR studies.

Weil and Votta used the CR method in the same AT&T environment but added an additional model and estimator - the Jackknife (JK) estimator for the M_h model, and compared it with the M_t - MLE estimator. They found that both estimators produced inaccurate estimates when their assumptions were violated. They also proposed a grouping method to improve these estimators but found that it only improved the accuracy of the M_t -MLE estimator [15].

Briand, et al., reported the first evaluation study that included one or two estimators from each of the four CR models. Using requirement artifacts inspected by NASA professionals, this study investigated the effect that the number of reviewers and the number of faults had on the estimates. The major results from this study showed that the CR models generally underestimate and recommended M h -JK as the best estimator. The results also

showed that the accuracy of the estimators improves with more inspectors and faults, finding that a minimum of four inspectors and six faults are needed to achieve satisfactory estimates. There was no improvement in accuracy beyond four inspectors and six faults [2]. Our current work builds on these early findings. A limitation of Briand, et al.'s, study was that their recommendations were based on only six inspectors using artifacts seeded with fifteen to twenty faults. Therefore, this study builds on their efforts to do a more detailed investigation using artifacts with real fault data and bigger data set. Similarly, Emam, et al., evaluated the CR estimators using only two inspectors and found M h to be the best CR model. They also advocated the use of subjective opinion with the CR estimates to make decisions on the need for re-inspection during real development [6, 7].

Therefore, most of the CR studies have utilized relatively small data sets. Up to that point, most of the estimation models assumed that the inspectors work individually. Then, a study was reported in which the inspectors collaborated with each other. This situation violated the assumption that inspectors work separately and therefore required the introduction of a new estimator. This new estimator was only compared with M t -MLE and produced similar results [9]. However, it is unknown how this estimator performs relative to the other estimators. Also, the advent of more effective inspection techniques like Perspective-Based Reading (PBR), which encourage team members to focus on different types of defects during their inspection thereby reducing the potential overlap, seems to directly contradict a basic requirement of the capture-recapture models. But, studies in this area revealed that capture-recapture models can also be applied to inspection techniques like PBR and yet again, M h -JK is the best estimator [20].

Most empirical studies in software engineering have evaluated the use of CR models on

software artifacts with a known number of seeded defects [2, 8, 9, 10, 11, 13, 16, 17-18, 20-21, 23-24]. However, in live software development, the actual defect count of an artifact is unknown. So, it is not clear what effect the use of seeded defects had on the estimation results. There is little evidence to support the efficacy of using CR models in real software development (with an unknown number of naturally occurring defects). We performed another study with the goal of evaluating the ability of the CR estimators to estimate the fault count of artifacts containing faults made during their development (as opposed to seeded faults). Each artifact was inspected twice, which allowed the analysis of the CR estimator's ability to decide about the need for re-inspection. The results showed that the estimates after second inspection were more accurate than the estimates after first inspection, and the CR estimates were accurate in determining the need of re-inspection after each inspection cycle [13].

The major results from the analysis of 10 years of research on the use of CR in software inspection as summed up by Petersson, et al. [9], and additional results from Walia, et al. [12, 13], are: a) CR models generally underestimate the fault count; b) M h - JK is the most accurate estimator when using data from four or more reviewers, c) the CR estimates improve with more input data, but there has not been much investigation of the effect of the number of inspectors and the number of faults on the performance of the CR models. While the researchers believe that using data from more reviewers as input to the models produces more accurate estimates, there is no clear consensus on the minimum number of reviewer's needed to obtain a satisfactory estimate. Some researchers recommend that data from a minimum of four reviewers is needed to obtain enough overlap for the models to be useful while other researchers recommend that data is needed from only two reviewers.

4. STUDY DESIGN

Previous empirical studies of Capture-Recapture (CR) in software inspections have evaluated the ability of the estimators to accurately predict the need of a re-inspection. The common finding from these evaluation studies is that the CR models generally underestimate the true fault count, but accuracy improves with more number of inspectors (or captures). The impact of the inspection team size on the estimation accuracy is expected to be positively correlated. However, this relation has not been empirically investigated.

Briand et al., reported the first comprehensive evaluation of the capture-recapture estimators on software requirement artifacts with a known number of seeded defects. They analyzed the impact that the number of inspectors had on the performance of estimators and recommended that four was the minimum number of inspectors required to obtain satisfactory estimates. However, this recommendation was based on a data set that had only six inspectors. Therefore, we believe that result is premature and is worthy of further study. To that end, our preliminary investigation in this area has provided positive evidence to motivate our hypothesis that the accuracy of CR estimators is positively correlated with the inspection team size [27]. This paper extends our initial work by incorporating a variety of different inspection data sets, and performing a comprehensive evaluation of the effect of the number of inspectors on the performance of CR estimators across the data sets with varying number of inspection team size. The selection of these data sets was guided by the following reasons.

To make a better comparison of the use of capture-recapture models in software engineering to their use in biology and wildlife research, we evaluated the CR models on a larger data set (more comparable in size to the data sets from biology or wildlife research). To address this need, one of the data set used in this study consisted of inspection data from seventy-

three inspectors. Furthermore, the data used in most of the previous CR evaluation studies was drawn from the inspection of artifacts with seeded faults, rather than naturally occurring faults. Therefore, this paper also investigated the effect of the number of inspectors on the CR estimates using five additional artifacts that were based on real system, and contained naturally occurring defects and were inspected by varying number of subjects (ranging from 6 inspectors to 8 inspectors to 17 inspectors).

These data sets were analyzed to re-evaluate the CR estimators used in previous studies by other researchers and to evaluate the estimators that have not previously been evaluated in the context of software inspections (i.e., the CMLE estimator for M_o and M_t type models, the EE estimator for all type of models, and the SC estimator for the M_h and M_{th} type of models). The findings from this study are then compared with the earlier findings in software engineering and with the findings from biology and wildlife research to gain useful insights about the applicability of these estimators for software inspections.

4.1 Research Goals

The main goal of this study is to understand how the performance of the estimators improves when increasing the number of inspectors. Stated more formally, the first goal of this study is to:

Analyze the capture-recapture estimators
For the purpose of characterizing the impact of the number of estimators
With respect to defect estimation accuracy and precision
From the point of view of the project managers and software developers.

The secondary goal of this study is to compare the relative performance of capture-recapture models and corresponding estimators shown in Table 2 with the increasing number of inspectors. Stated more formally, the second goal is to:

Analyze the capture-recapture models and estimators
For the purpose of evaluation
With respect to defect estimation accuracy and precision
From the point of view of software organizations.

4.2 Data Set

The data for capture-recapture analysis in this paper includes six different inspection data sets (with varying number of inspection team size). These data sets are shown in Table 3 and were drawn from earlier inspection studies conducted at Mississippi State University (MSU) and an inspection study conducted at Microsoft Research.

Table 3 - Capture-Recapture Data Sets

Data Set	Artifact Name	Description	Number of Inspectors	First Inspection Defects	Total Number of Defects
1	Loan Arranger Financial System	Grouping loans into bundles based on user-specified characteristics	73	NA	30
2	Starkville Theatre System	Management of ticket sales and seat assignments for the community theatre	8	30	55
3	Management of Apartment and Town properties	Managing apartment and town property, assignment of tenants, rent collection, and locating property by potential renters	8	46	105
4	Conference Management	Helping the conference chair to manage paper submission, notification of results to authors, and other related responsibilities	6	52	94
5	Conference Management	Same as Above	6	64	118
6	Data Warehouse Functional Requirements	The functional, and other (e.g., security, performance, interface) requirements of Data Warehouse	17	169	253

The details and findings of the original studies have been published [28-29]. Only the information that is relevant to the capture-recapture analysis is provided in this section. The data sets are grouped (based on the similarity in the nature of the artifacts inspected, defects found, and the inspectors employed) and described in the following three subsections.

4.2.1 Data Set 1

Data Set 1 was drawn from an earlier inspection study that was conducted at Microsoft Research to investigate the impact of educational background on the effectiveness of an inspector.

Artifacts: The artifact inspected during this study was a generic (i.e., non-Microsoft) requirements document describing the requirements for the Loan Arranger financial system. The Loan Arranger system is responsible for grouping loans into bundles based on user-specified characteristics. These loan bundles are then sold to other financial institutions.

Defects: For use in previous studies [27], the document was seeded with thirty realistic defects. The defect seeding was done by researchers other than the authors of this paper prior to the design of the capture-recapture study. Therefore the defects that were seeded should not provide any bias in the current study.

Inspectors: There were seventy-three (73) inspectors who were drawn from an internal training course taught by the Microsoft Engineering Excellence group. One of the main goals of the course was to teach participants about inspections and their use at Microsoft. The participants were drawn from all major product groups across Microsoft. About 70% had bachelor's degrees with the other 30% having Master's degrees. On average, the participants had about 2 years of experience working in the field

Inspection Process: First, the participants received training on the basic concepts involved in an inspection process. Then, the participants performed their own inspection of the Loan Arranger requirements document. To guide their review of the document, the participants used a standard fault-checklist. During the inspection, each participant worked alone to identify and record as many defects as possible. They were given seventy minutes to complete the

inspection task. At the conclusion of the inspection, the seventy-three individual defect lists were collected and processed. The processing involved determining which of the thirty seeded defects were found by each participant. It is this information that was used as raw data for the capture-recapture study described in the remainder of this paper.

4.2.2 Data Sets 2, 3, 4, and 5

Data Set 2, 3, 4, and 5 were drawn from earlier inspection studies conducted at Mississippi State University (MSU). The original goal of these studies was to investigate the impact of errors (i.e., mistakes) committed during the development of the requirement document [28-29].

Artifacts: The artifacts used in these data sets were real requirement documents. These artifacts were developed by senior-level undergraduate students, majoring in either computer science or software engineering enrolled in the Software Engineering Senior Design Course at MSU during the Fall 2005 and Fall 2006 semesters. The sixteen subjects in Fall 2005 semester were divided into two 8-person teams that developed the requirement document for their respective system (i.e., *Starkville Theatre System* and *Management of Apartment and Town Properties*) as shown in Table 3. Similarly, twelve subjects in Fall 2006 semester were divided into two 6-person teams that developed separate requirement document for the *Conference Management* system. The course required student teams to interact with real customers, elicit, and document requirements that they would later implement. So, even though the developers are students, the artifacts are realistic for a small project. A brief description of the requirement artifacts belonging to each of these four data sets is provided in Table 3.

Defects: Unlike the artifact used in Data Set 1 (that was seeded with realistic defects), the artifacts used in the data sets 2 through 5 included natural defects that were made by developers during the development of these artifacts.

Inspectors: Each artifact was inspected for defects by the same set of developer's who created these artifacts. The number of inspectors for each artifact is also shown in Table 3.

Inspection Process: Each artifact was inspected twice by the same inspectors. During the first inspection, the subjects received training on a fault checklist. Then, each inspector individually inspected the artifact using the fault checklist and logged any faults identified. After the first inspection, inspectors met as a team to consolidate their faults into a team fault list for each artifact. During the second inspection, the subjects were trained on how to abstract errors from faults, how to classify the errors, and how to use the errors to re-inspect the requirements document. Then, each inspector re-inspected the artifact using the errors to find the additional faults. The same inspection process was followed by the subjects in each team, and the artifacts were not modified or corrected between inspections (i.e., the same artifact was re-inspected). The number of faults found during the first inspection and the total number of faults found after two inspections in each artifact is shown in the last two columns of Table 3. For example, for Artifact used in data set 2 (i.e., *Starkville Theatre System*), 8 inspectors found 30 distinct faults during the first inspection, and found another 25 new faults during the second inspection totaling the fault count at 55 (as shown in the last column).

For the purpose of the evaluation in this study, only the data from the first inspection is used for calculating the capture-recapture estimates because the CR estimators requires input data from individual inspections for it to produce an estimate. The data from the second inspection is only used to calculate the total number of faults that is assumed to be the actual

fault count of an artifact for the sake of the evaluation. Using the data only from first inspection also help us control the variability of the inspection technique employed, since all four data sets use the same inspection technique (i.e., fault checklist) during the first inspection.

4.2.3 Data Sets 6

Data Set 6 was also drawn from another inspection study conducted at Mississippi State University (MSU).

Artifacts: The artifact inspected during this study was a natural language requirements specification document for a data warehouse system that was developed by professional developers at the Naval Oceanographic Office. The document was 30 pages long and included the overview (scope and purpose of the system), the functional requirements, and other (e.g., security, performance, interface) requirements.

Defects: Like Data Sets 2-5, this data set also included natural defects that were made by developers during the development of the artifact.

Inspectors: A total of 18 graduate students enrolled in the Software Verification and Validation (V&V) course or the Empirical Software Engineering (ESE) course at MSU inspected the requirement document. These participants did not develop the requirements document, nor did they have access to any of the developers of the requirement document.

Inspection Process: Similar to the data sets 1 through 5, each participant inspected the artifact twice. The first inspection data (during which subjects individually used the fault checklist method to log defects) is used as input to the capture-recapture analysis and the number of unique faults found at the end of two inspection cycle is assumed to be the total fault count.

4.3 Evaluation Procedure

To compare the performance of the estimators using data from a varying number of

inspectors as input, virtual inspection teams were created for each inspection team size (e.g., for Data set 1, we varied the inspection team size ranging from one inspector to seventy-three inspectors). The process of creating virtual inspection teams consisted of randomly selecting the appropriate number of inspectors from the overall pool of inspectors. For example, to create the fifteen member inspection teams in data set 1, fifteen inspectors were randomly selected. Then, a matrix of the inspection data (containing 15 columns and 30 rows) from these fifteen inspectors was created by keeping the fault count constant. Using this approach, 100 virtual inspection teams were created for each team size, i.e. 100 virtual inspection teams of size two, another 100 virtual inspection teams of size three, and so on, up to a team size of seventy-three. This process resulted in the creation of 100 inspection teams (if possible) for each inspection team size (1- 72) and one team that combines all the seventy-three inspectors. Similar process for varying the inspection team size was performed for all the other data sets shown in Table 3.

An automated script developed by the researcher was used to generate the 100 possible sub-matrices for each inspection team size for all the input data sets. The script then fed these input data sets to the automated tool CARE-2 [7], (originally developed for the biology and wildlife research) in order to calculate the capture-recapture estimates. So, executing the script that interacted with the CARE-2 tool, the appropriate matrices were created for each inspection team size and produced the CR estimates of the total number of defects for all the different data sets and all the CR estimators.

4.4 Evaluation Criteria

For each inspection team size (i.e.,, 1-73 for data set 1, 1-8 for data sets 2 and 3, 1-6 for data sets 4 and 5, and 1-17 for data set 6), the hundred possible estimates produced are used to compute the median estimate. The estimators are then evaluated on their performance using three

parameters: accuracy (bias), precision (variability), and failure rate. These metrics are explained below and are illustrated in Figure 2 with example of an inspection team size of two inspectors.

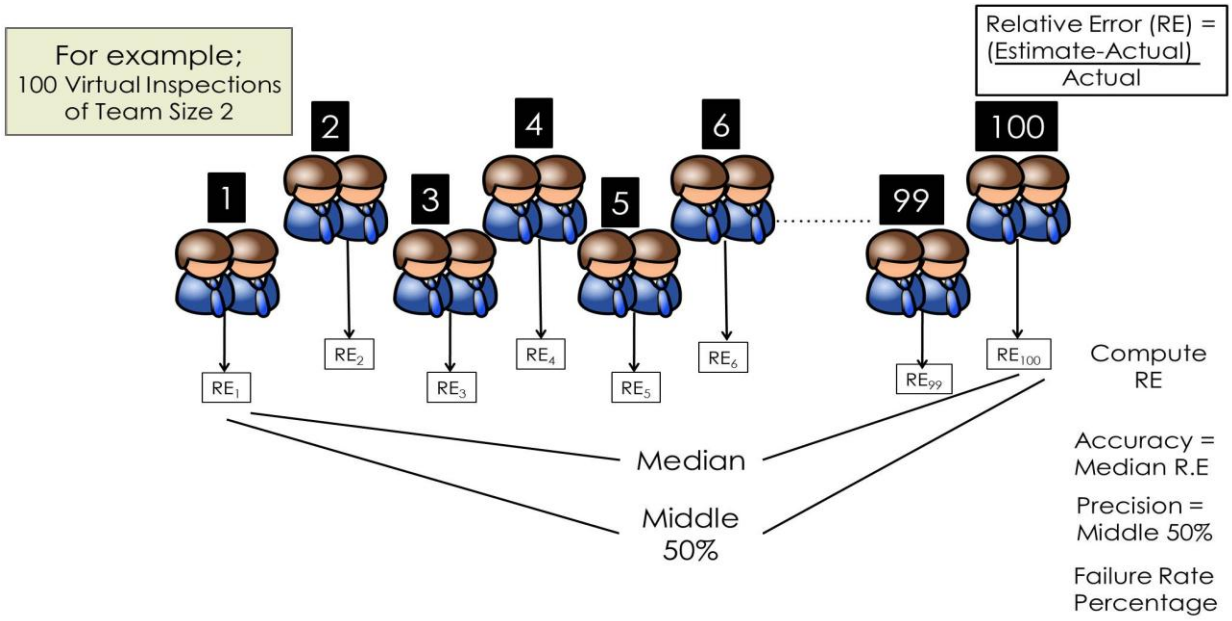

Figure 2 - Evaluation Criteria for CR Estimators: Accuracy, Precision, and Failure Rate

The **accuracy (bias)** is measured as the relative error (R.E) of an estimate. It is calculated as:

Relative error = (Estimated number of defects – Actual number of defects)/ Actual number of defects

A R.E of zero means absolute accuracy (zero bias), a positive R.E. means an overestimation, and a negative R.E means an underestimation. The accuracy of the estimator is measured by calculating the median relative error for each inspection team size. According to Eick, et al. and Briand, et al., the accuracy of an estimate is considered satisfactory when the R.E. is within +/- 20% of the actual value [3, 10]. In this paper, we have evaluated the accuracy of estimates at varying levels of R.E (e.g., +/- 20%, +/- 10%, +/- 5%, 0% etc.).

Because we do not know the actual number of defects for data sets 2 through 6, the total number of exclusive defects found after both inspections is assumed to be the actual defect count for the purposes of this study. The difference between the estimated defect count and this actual defect count is used to evaluate the accuracy of CR estimators. Furthermore, the error in the estimates is calculated relative to each artifact to allow for combination of the results from all the artifacts.

The **precision** of an estimator is measured by calculating the variability of the R.E. estimates for each input size (e.g., 1-73). R.E variability around the central tendency i.e. (median value) is measured using the inter quartile range of the 25^{th} percentile to the 75^{th} percentile.

The **failure rate** of an estimator is defined as the number of time an estimator fails to produce any result. Because each estimator makes different assumptions about the data and they all operate on the same data matrix, some estimators can fail if the actual data fails to meet some of its basic assumptions.

5. ANALYSIS AND RESULTS

This section provides analysis of the capture-recapture estimates and is organized around the two research goals described in Section 4.1. Rather than discussing each data set in chronological order, to reduce duplication we have grouped the results based on data type. Section 5.1 evaluates the CR estimators using data set 1 which contain artifact that was 30 seeded defects. Section 5.2 discusses results from data sets 2 through 5 which contain natural occurring defects made by student teams during the artifact development. Finally, Section 5.3 discusses results from data set 6 that uses requirement artifact developed by professional developers and inspected by student inspectors for real naturally occurring defects.

5.1 Evaluation of Capture-Recapture Models and Estimators on Data Set 1 (Microsoft)

Our main research goal deals with evaluating the effect of inspection team size on the estimates produced by capture recapture models. To provide an overview of the result, Figure 3 shows the median relative error for each capture-recapture estimator across all team sizes for Data Set 1 (with inspection team size varying from one through seventy-three). All the estimates with the same estimator are connected with a line. The result in Figure 3 shows that the CR estimators severely underestimate the actual fault count (i.e., with a relative error in excess of -30%) when the number of inspectors is small (i.e., 1 through 10); and the CR estimators shows a consistent improvement in their accuracy with more number of inspectors. The shaded lines in Figure 3 show the region of +/-20% within which the estimate produced by the CR estimators is considered satisfactory [3, 10]. The result in Figure 3 also reveals that the majority of CR estimators need a minimum of 17 inspectors to achieve a median estimate of relative error within +/-20%. The result also showed that some of the CR estimators improved faster (i.e., obtained

median estimates within +/-20% with fewer number of inspectors) as compared to other estimators.

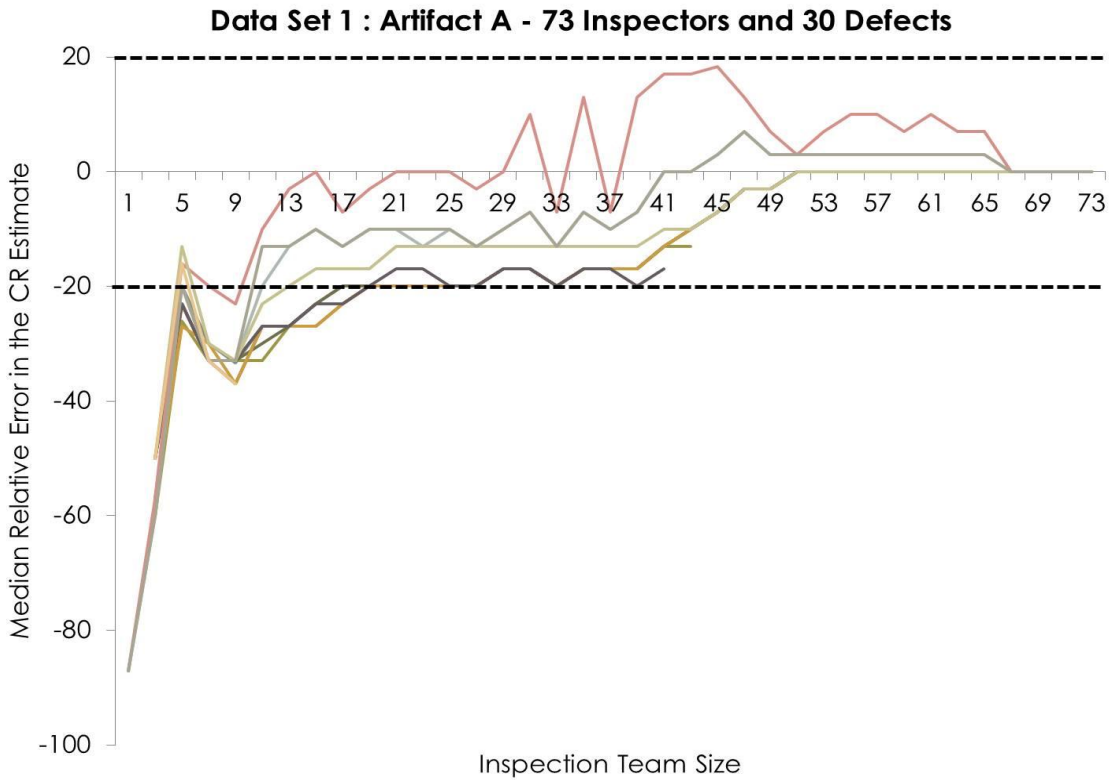

Figure 3 - Median Relative Error in the Estimates vs. Inspection Team Size for Data Set 1

Our second research question deals with characterizing the relative improvement in the performance of the different CR estimators with respect to varying level of relative error (R.E) in the estimate. To that end, Table 4 provides insights into the number of inspectors required by each CR estimator to obtain an estimate within 0%, +/- 5%, and so on up to +/- 40% of the actual fault count. The inspector count for each R.E percentage shown in Table 4 is calculated so that beyond that point, the median estimate is always less than the given R.E percentage (i.e., from that point forward, the R.E decreases as the inspection team size increases). For example, for Mo-CMLE estimator, 10 inspectors are required to achieve a median estimate with a relative

error less than -30%, and from 11 inspectors and beyond, the estimate is always less than -30% of the actual value.

Table 4 - Number of Inspectors Required to achieve Different Levels of Estimation Accuracy for CR Estimators

Estimators	0%	-5%	-10%	-15%	-20%	-25%	-30%	-35%	-40%
M_o-CMLE	50 (0.3)	46 (0.3)	42 (0.3)	40 (0.5)	**16 (0.8)**	14 (2)	10 (3)	4 (26)	4 (26)
M_o-UMLE	50 (0.3)	46 (0.3)	42 (0.3)	40 (0.5)	**18 (0.8)**	16 (0.7)	10 (7)	10 (7)	4 (23)
M_o-EE	Failed			40 (0.3)	**18 (0.8)**	14 (0.7)	12 (0.5)	4 (26)	4 (26)
M_t-CMLE	50 (0.3)	46 (0.3)	42 (0.3)	40 (0.5)	**16 (0.8)**	14 (1)	10 (3)	4 (25)	4 (25)
M_t-UMLE	50 (0.3)	46 (0.3)	42 (0.3)	40 (0.5)	**18 (0.8)**	16 (1)	10 (3)	4 (22)	4 (22)
M_t-EE	Failed				**18(0.8)**	14 (1)	10 (3)	4 (26)	4 (26)
M_h-SC	66 (0.1)	48 (1)	34 (4)	12 (2)	**10 (11)**	10 (11)	10 (11)	4 (84)	4 (84)
Mh-JK	66(3)	66 (0)	48 (4)	46 (3)	**10 (20)**	4 (29)	4 (29)	4 (29)	4 (29)
M_h-EE	50 (0.3)	46 (0.3)	40 (0.1)	20 (0.3)	**12 (0.3)**	10 (3)	10 (3)	4 (39)	4 (39)
M_{th}-SC	66 (0.1)	40 (1)	34 (4)	10 (3)	**10 (3)**	10 (3)	10 (12)	4 (93)	4 (93)
M_{th}-EE	Failed								4 (44)

Based on the results shown in Table 4, some general observations are as follows:

a) Combining the results from all the CR estimators, depending on the type of estimator, somewhere between 10 to 18 inspectors are required to achieve a satisfactory estimate (i.e., within relative error of +/- 20%);

b) Estimators corresponding to M_h model and M_{th} model obtain an estimate within 20% of the actual value with fewer number of inspectors (10 or 12 inspectors) compared with the estimators for models M_o and M_t (16 or 18 inspectors);

c) EE estimators for all the models (M_o-EE, M_t-EE, M_h-EE, and M_{th}-EE) exhibit failure rate even for the larger number of inspectors. Among all the EE estimators, M_h-EE exhibits lowest failure rate. The other EE estimators (M_t-EE and M_{th}-EE) more often failed to produce the defect estimate.

Based on these results in Table 4, the accuracy of Jackknife (M_h-JK) estimator and the SC estimators (for M_h and M_{th} models) improves faster with increasing inspection team size as compared to the other CR estimators. Therefore, based on the median R.E values, JK and SC are the best estimators.

While the above results demonstrate the accuracy of CR estimators, we also examined the variance in the estimates (across an array of 100 estimates) at each inspection team size for all the CR estimators to gain insights into the relative *precision* and *reliability* of an estimator. Table 4 provides these variance values for different inspector counts in parenthesis. It is calculated as the size of the interquartile range (i.e., the spread of the middle 50% of the estimate data). The variance values at selected inspection counts in Table 4, reveals that the M_h-SC, M_h-JK and M_{th}-SC estimators (which required fewer number of inspectors to achieve accurate estimates) shows a higher degree of variability in their estimates as compared to the other estimators. For example, M_h-JK estimator only required 10 inspectors to achieve an estimate within +/- 20% of actual value as compared to the CMLE and UMLE estimators (that needed 16 or 18 inspectors), but the variability in the estimates produced from the JK estimator is twenty-five folds (20/0.8) the variability in the estimates produced by CMLE and UMLE estimators.

To properly understand the trends in the precision of an estimator with increasing inspection team size, the variability values for each estimator at each team size is shown in Figure 4. Some general observations from Figure 4 are that:

a) The CR estimators show high variability (i.e., *lack of precision*) in the estimates with small number of inspectors, but the variability values show a consistent decrease as the number of inspectors increase; and the CR estimators belonging to M_o, and M_t models

become precise faster (i.e., with fewer number of inspectors) as compared to the estimators belonging to M_h and M_{th} models;

b) The CR estimators for M_h and M_{th} models show a higher and inconsistent decrease in the variability values compared to the estimators for M_o and M_t models. *For example*, the estimates obtained from the Jackknife estimator (as shown in shaded line in Figure 4) shows an sudden increase in the variability even with larger number of inspectors (around 37 to 47 inspectors) in comparison to other CR estimators.

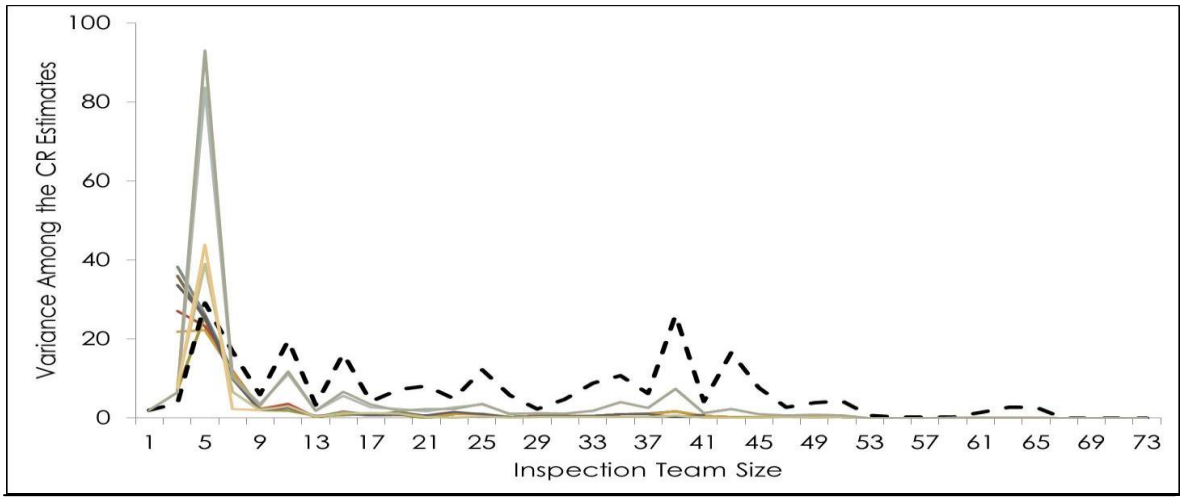

Figure 4 - Variability in the Estimates vs. Inspection Team Size for Data Set 1

Therefore, to evaluate the reliability of the CR estimators, we need to analyze both the accuracy (i.e, R.E) and the precision (i.e., R.E. variability) as the number of inspector increases. An approach for combining the analysis of accuracy and precision of an estimate is to calculate the three different values for each inspection team size (from 1 -73) and for each CR estimator combination. These following three values are calculated from an array of 100 estimates:

a) The *median* estimate (50^{th} percentile),

b) The *seventy-fifth largest* estimate (75^{th} percentile), and

c) The *twenty-fifth largest* estimate (25th percentile).

Together b) and c) define the interquartile range and is essentially the range of the middle 50% of the estimates. Figure 5 shows the relative error in the estimate at all these three values with relative errors (R.E) in the median estimate appearing between the upper (75th percentile) and lower bound (25th percentile) on the estimate. The result for all the estimators is shown in Figure 5, except the M_o-EE and M_t-EE estimators because of their high failure rate.

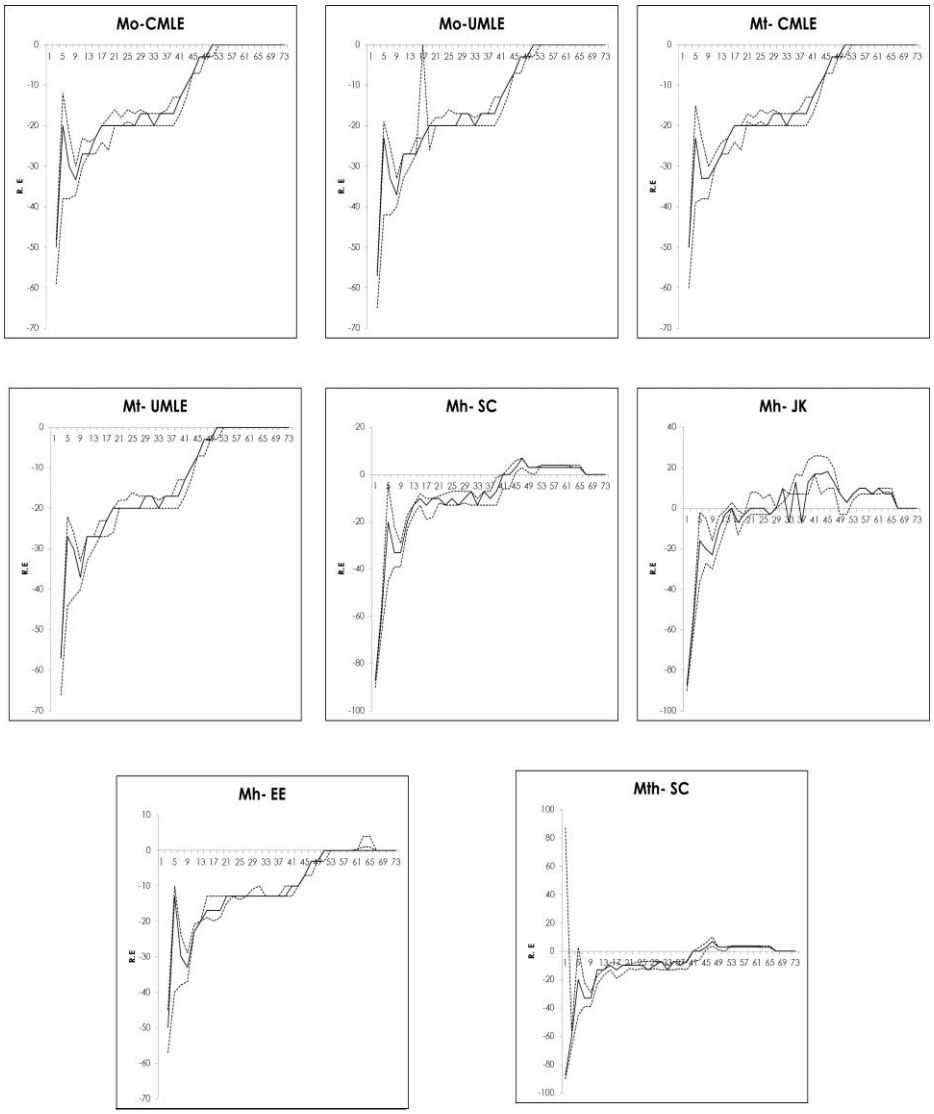

Figure 5 - Combining the Results from Accuracy and Precision of an Estimator vs. Inspection Team Size for Data Set 1

Figure 5 indicates that the number of inspectors directly influence the accuracy as well as precision of the CR estimators. To quantify these results, we wanted to determine the cut-off points.

Similar to the results shown in Table 4 (that was only based on the median R.E values), we analyzed the *median* estimate, the *seventy-fifth largest* estimate, and the *twenty-fifth largest* estimate values (as shown in Figure 5) to determine how many inspectors are required to achieve an estimate at varying levels of estimation accuracy and precision (e.g., +/- 30%, +/- 20%, +/- 10%, +/- 0% etc.). This analysis was performed separately for all the CR estimators and is described in detail for Jackknife estimator in this section and illustrated in Figure 6.

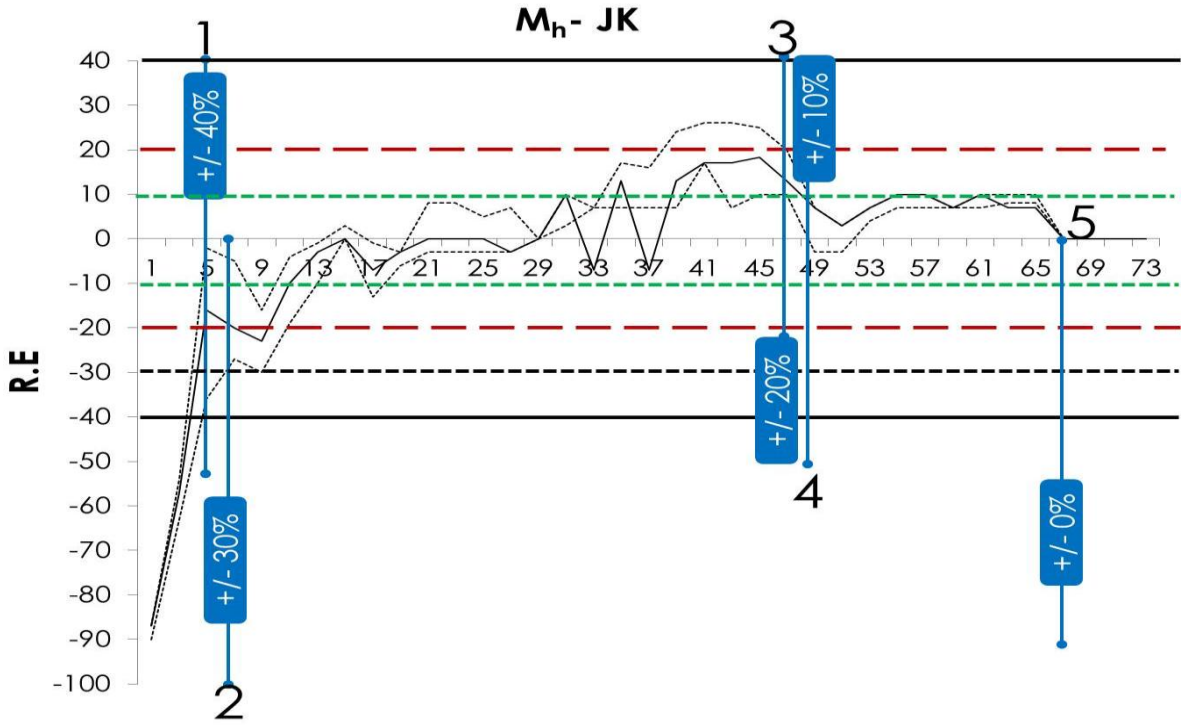

Figure 6 - Determining Cut-off points for the Jackknife Estimator

Figure 6 shows the number of inspectors (shown by solid vertical lines) required to achieve an estimate at +/- 40%, +/- 30%, +/- 20%, +/- 10% and at +/- 0%. The process of determining the cutoff points for these regions is described as follows:

a) *Cutoff at +/-40%*: The first vertical line (labeled 1) in Figure 6 shows that the JK estimator requires 5 inspectors where all the three values (median, 75th percentile, and 25th percentile) have a relative error less than or equal to 40% and they never exceed 40% as the number of inspector increases.

b) *Cutoff at +/-30%*: The second vertical line (labeled 2) in Figure 6 shows that the JK estimator requires 7 inspectors where all the three values (median, 75th percentile, and 25th percentile) have a relative error less than or equal to 30% and they never exceed 30% as the number of inspector increases.

c) *Cutoff at +/-20%*: The third vertical line shows that the JK estimator requires 46 inspectors where all the three estimates had a relative error less than or equal to +/-20% and they never exceed +/- 20% as the number of inspector increases.

d) *Cutoff at +/-10%*: The fourth vertical line shows that the JK estimator requires 48 inspectors where all the three estimates had a relative error less than or equal to +/-10% and they never exceed +/- 10% as the number of inspector increases.

e) *Cutoff at +/-0%*: The first vertical shows that the JK estimator requires 66 inspectors to achieve absolute accuracy and precision.

The process for determining the cutoff values (i.e., the number of inspectors) was same for all the other CR estimators. The resulting output of this process is shown in Table 5. The left side of Table 5 shows the minimum number of inspectors required to achieve varying levels of estimation accuracy and precision for all the CR estimators (based on the process shown in Figure 6). The right side of Table 5 shows the minimum number of inspectors required to achieve a median estimate within +/- 20% range (based on the results shown in Table 4).

Table 5 - Number of Inspectors Required for Achieving Different Levels of Estimation Accuracy and Precision for CR Estimators vs. Results from Table 4

Accuracy + Precision + Failure Rate

Estimators	0%	-5%	-10%	-15%	-20%	-25%	-30%	-35%	-40%
Mo-CMLE	52	48	44	42	***20***	20	11	11	5
Mo-UMLE	52	48	44	42	***20***	20	14	11	10
Mo-EE	*Exhibits failure rate beyond inspection team size of 43*			42	***20***	20	11	10	9
Mt-CMLE	52	48	44	42	***20***	16	11	10	5
Mt-UMLE	52	48	44	42	***20***	20	13	10	9
Mt-EE	64	64	64	64	***20***	20	11	10	5
Mh-SC	66	48	40	20	***12***	10	10	10	7
Mh-JK	66	66	48	48	***46***	46	10	7	5
Mh-EE	66	48	42	21	***13***	10	10	10	5
Mth-SC	66	48	36	20	***12***	11	10	10	6
Mth-EE	*Exhibits failure rate beyond inspection team size of 10 and higher*								9

Accuracy

Estimators	-20%
Mo-CMLE	***16***
Mo-UMLE	***18***
Mo-EE	***18***
Mt-CMLE	***16***
Mt-UMLE	***18***
Mt-EE	***18***
Mh-SC	***10***
Mh-JK	***10***
Mh-EE	***12***
Mth-SC	***10***
Mth-EE	

For example, the first row of the left side of Table 5 shows that the Mo-CMLE estimator needs 5 inspectors to achieve all three estimates (median, 75th percentile, and 25th percentile) less than or equal to +/- 40%, 11 inspectors to get into +/- 30% range, 20 inspectors to get into +/- 20% range, and so on. Whereas, the first row of the right hand side of Table 5 shows that the M_0-CMLE estimator needs 16 estimators to achieve an estimate within 20% range based only on the median estimate. The result from Table 5 shows that there is not a huge difference in the cutoff values at +/- 20% when considering the estimation and precision values vs. just the accuracy values of CR estimators, except in case of the Jackknife estimator. Regarding the JK estimator, it needs considerably larger number of inspectors (i.e., 46 vs. 10 inspectors) in order to achieve an accuracy and precision within +/- 20% range in comparison to the other CR estimators. This shows that the JK is an imprecise estimator (i.e., large variability) with fewer numbers of inspectors and needs a large number of inspectors to achieve a satisfactory estimate. Overall, the

major insights provided gained from the results provided in this section are summarized as follows:

a) There is a direct improvement in the accuracies (i.e., median R.E) and the precision (i.e., interquartile range) as the number of inspector increases.

b) The minimum number of inspectors needed to achieve a satisfactory estimate (i.e., within +/- 20%) varies from 12 to 20 depending on the estimator. Only the JK estimator needs in excess of 40 estimators to achieve a satisfactory estimate.

c) The estimators corresponding to M_h and M_{th} models (except the JK estimator) need fewer number of inspectors to achieve a satisfactory estimate (i.e., within +/- 20%) as compared to the Mo and Mt models.

d) The EE estimator for M_o model exhibits failure to produce an estimate for all of the possible 100 combinations beyond 42 inspectors. The EE estimator for M_h model also failed to produce an estimate for some (but not all) of the virtual inspections. Similarly, the M_{th}-EE estimator is also not recommended because of its inability to produce an estimate for inspection team size of 10 or more inspectors.

e) Considering the accuracy and precision values, the JK estimator is not recommended because of the huge variability among the estimates with less number of inspectors. This recommendation is contradictory to the findings by previous researchers who have recommended the JK estimator to be most accurate estimator based on the median estimate values alone.

f) Finally, Sample Coverage (SC) estimators for M_h and M_{th} models are recommended to be the best CR estimators for use with twelve inspectors. This is a new result since the SC estimators have not been previously studied in the software engineering research.

5.2 Evaluation of Capture-Recapture Models and Estimators on Data Sets 2, 3, 4 and 5

The results in Section 5.1 were based on the inspection of a requirement document that was seeded with defects prior to the inspection. Likewise, researchers in the past have always evaluated the use of CR on software artifacts with seeded faults. To that end, software reliability research has also shown that seeded, artificial defects differ in detection probability from naturally occurring defects and are easier to detect. Even while re-seeding realistic defects, their densities differ from that of natural occurring defects [14].

Therefore, the nature of the defects can influence the estimation results. As in live software development, the actual defect count of an artifact is unknown after an inspection. To provide better information for project managers on the *number of inspectors* to use when deciding on the adoption of CR in their organizations, it is imperative to evaluate the effect of the number of inspectors on the CR estimates in real settings.

This section evaluated the effect of the *number of inspectors* on the CR estimates using data from inspection of four different real software artifacts that were developed by students in senior-level capstone software engineering class (i.e. they were created to guide the later implementation of the system). These artifacts contained naturally occurring defects that were committed by student teams during the development, and were later inspected in the same environment. In addition, each artifact was inspected twice, which allow us to count the total number of exclusive defects found after two inspections and is assumed to be the actual defect count for the purposes of this study.

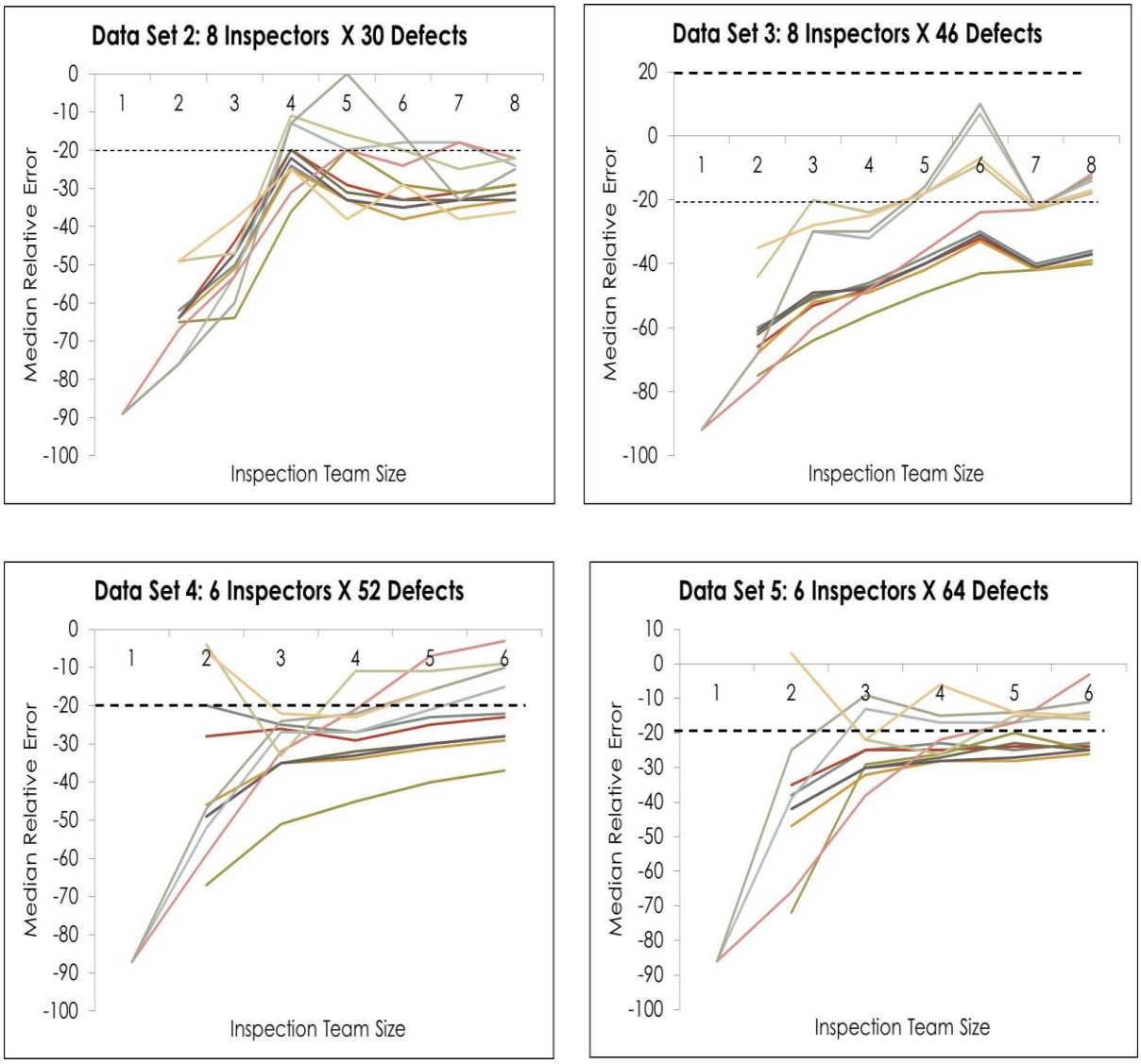

Figure 7 - Median Relative Error in the Estimates for Data Sets 2, 3, 4 and 5

To mirror the live software development settings, the data from first inspection is used to evaluate the accuracy and precision of CR estimators for the reasons mentioned in Section 4.2.2. Also, as mentioned earlier, the same inspection technique (i.e., *Fault Checklist*) was used to inspect all the four artifacts during the first inspection. The difference between the estimated defect count (*using data from the faults found during the first inspection*) and the actual defect count (*i.e., the total number of exclusive faults found at the end of two inspection cycles*) is used

to evaluate the accuracy and precision of CR estimators. Furthermore, the error in the estimates is calculated relative to each artifact to allow for combination of the results from all the four artifacts. To provide an overview of the results, Figure 7 shows the median relative error for each CR estimator across all team sizes for Data Sets 2 and 3 (with inspection team size varying from 1 to 8), and for Data Sets 4 and 5 (with inspection team size varying from 1 to 6). From this Figure, some interesting observations can be made as discussed follows:

a) Regarding Data set 2 (where 8 inspectors found 30 exclusive defects during the 1st inspection cycle), the results confirm the results (from Section 2.1) that the CR estimators severely underestimate the actual fault count with eight inspectors and less. The median estimates for the CR estimators in Data set 2 never stay within -20% relative error as the inspection team size increases.

b) Regarding Data set 3, the estimation results are somewhat better as compared to the results obtained from Data set 2:

    a. The median estimate for the Mh and M_{th} models is within -20% at inspection team size of eight inspectors. However, there is not a consistent improvement in the median estimate with increase in inspection team size. For example, for SC estimator for Mh model, the relative error in the median estimate goes from -18% with 5 inspectors to +7% with 6 inspectors to -23% with 7 inspectors and -14% with 8 inspectors. Therefore, it is hard to evaluate if the median estimate would have stayed within or fell outside -20% range with an inspection team size of more than eight inspectors.

c) Regarding Data sets 4 and 5, similar trend was noticed

a. The median estimates from M_0 and M_t models severely underestimate (in excess of -30%) the actual fault count with six inspectors as less. On the other hand, all the CR estimators belonging to M_h and M_{th} models produce an estimate within +/- 20% of the actual count for inspection team sizes of 5 and 6 inspectors.

b. Again, there is an inconsistent improvement in the median estimate with increasing inspection team size. This result also highlights the threat of using data sets of smaller number of inspectors (only six inspectors used in data sets 4 and 5) to objectively evaluate the minimum number of inspectors required to achieve an accurate estimate.

Similar to the analysis process that combined the accuracy and precision for Data Set 1 (as shown in Section 5.1), we calculated the median, 25^{th} percentile, and 75^{th} percentile from an array of 100 estimates for each CR estimator at each inspection team size for all the four data sets (2, 3, 4 and 5) separately. We analyzed these three values (the median estimate, the seventy-75^{th} largest estimate, and the 25^{th} largest estimate) to determine the number of inspectors required to achieve an estimate at varying levels of estimation accuracy and precision (e.g., +/- 20%, +/- 10%). The process of determining the cut-off points is same as described in Section 5.1 (i.e., all three values have a relative error less than or equal to a certain level and they never exceed that level as the number of inspector increases). The result regarding the minimum number of inspectors for achieving varying levels of accuracy and precision for all the CR estimators is shown in Table 6. The result from Table 6 confirms some of the results obtained from Data Set 1, contradict some of the earlier results, and provide some additional insights as discussed below:

Table 6- Number of Inspectors Required for Achieving Different Levels of Estimation Accuracy and Precision for CR Estimators for Data Sets 2, 3, 4 and 5

Data Set 2

Estimators	-10%	-20%	-30%
M_o-CMLE	Never within this range		8
M_o-UMLE	Never within this range		
M_o-EE	Never within this range		8
M_t-CMLE	Never within this range		
M_t-UMLE	Never within this range		
M_t-EE	Never within this range		
M_h-SC	Never within this range		8
Mh-JK	Never within this range		6
M_h-EE	Never within this range		8
M_{th}-SC	Never within this range		8
M_{th}-EE	Never within this range		

Data Set 3

Estimators	-10%	-20%	-30%
M_o-CMLE	Never within this range		
M_o-UMLE	Never within this range		
M_o-EE	Never within this range		
M_t-CMLE	Never within this range		
M_t-UMLE	Never within this range		
M_t-EE	Never within this range		
M_h-SC	Never within this range	8	8
Mh-JK	Never within this range	8	6
M_h-EE	Never within this range	8	5
M_{th}-SC	Never within this range	8	5
M_{th}-EE	Never within this range	8	5

Data Set 4

Estimators	-10%	-20%	-30%
M_o-CMLE	Never within this range		5
M_o-UMLE	Never within this range		6
M_o-EE	Never within this range		
M_t-CMLE	Never within this range		6
M_t-UMLE	Never within this range		6
M_t-EE	Never within this range		6
M_h-SC	Never within this range	6	5
Mh-JK	Never within this range	6	5
M_h-EE	Never within this range	6	4
M_{th}-SC	Never within this range	6	4
M_{th}-EE	Doesn't produce estimate		5

Data Set 5

Estimators	-10%	-20%	-30%
M_o-CMLE	Never within this range		5
M_o-UMLE	Never within this range		4
M_o-EE	Never within this range		6
M_t-CMLE	Never within this range		6
M_t-UMLE	Never within this range		6
M_t-EE	Never within this range		6
M_h-SC	Never within this range	6	4
Mh-JK	Never within this range	6	4
M_h-EE	Never within this range	6	3
M_{th}-SC	Never within this range	6	4
M_{th}-EE	Never within this range	6	3

a) Confirmation results: These results confirms our findings from Data Set 1

   a. There is a direct improvement in the accuracy (i.e., median R.E) and the precision (i.e., interquartile range) as the number of inspector increases.

   b. The CR estimators belonging to M_o and M_t models require more than eight inspectors to achieve a satisfactory estimate..

   c. The EE estimators exhibit a high failure frequency Therefore, the EE estimators are not recommended because of their inability to produce an estimate for smaller number of inspectors.

d. The Sample Coverage (SC) estimators for M_h and M_{th} models are recommended for use with smaller number of inspectors.

b) Contradictory results: These results contradict our findings from Data Set 1

a. The CR estimator belonging to M_h and M_{th} models can achieve a satisfactory estimate with six to eight inspectors (depending on the data set being used). However, due to smaller number of inspectors and an inconsistent improvement in the estimation accuracy, we cannot completely recommend this result.

b. Considering the accuracy and precision values, the JK estimator is also recommended for use with smaller number of inspectors.

5.3 Evaluation of Capture-Recapture Models and Estimators on Data Set 6

While the results provided in Section 5.2 are based on the inspection data of the requirement documents that contained real faults, the artifacts inspected in those data sets were developed by student teams in senior-level capstone project, and it may not be representative of industrial strength requirement document. Also, students in a classroom setting are likely to have different experience and time pressures than would be of true professionals in a real environment, and may commit different defects during the development.

This section provides analysis of the CR estimates using data from an inspection of a natural language requirements document that was developed by software professionals at Naval Oceanographic Office and contained natural defects that were made during the development. The inspection data set (i.e., Data Set 6 in Table 3) used for this analysis is described in Section 4.2.3. As mentioned in Section 4.2.3, this document was inspected twice by seventeen subjects. Like

data sets 2 through 5, the data from first inspection is used for the CR analysis, and the total number of unique faults found at the end of two inspection cycles is assumed to be total fault count for the purpose of evaluation.

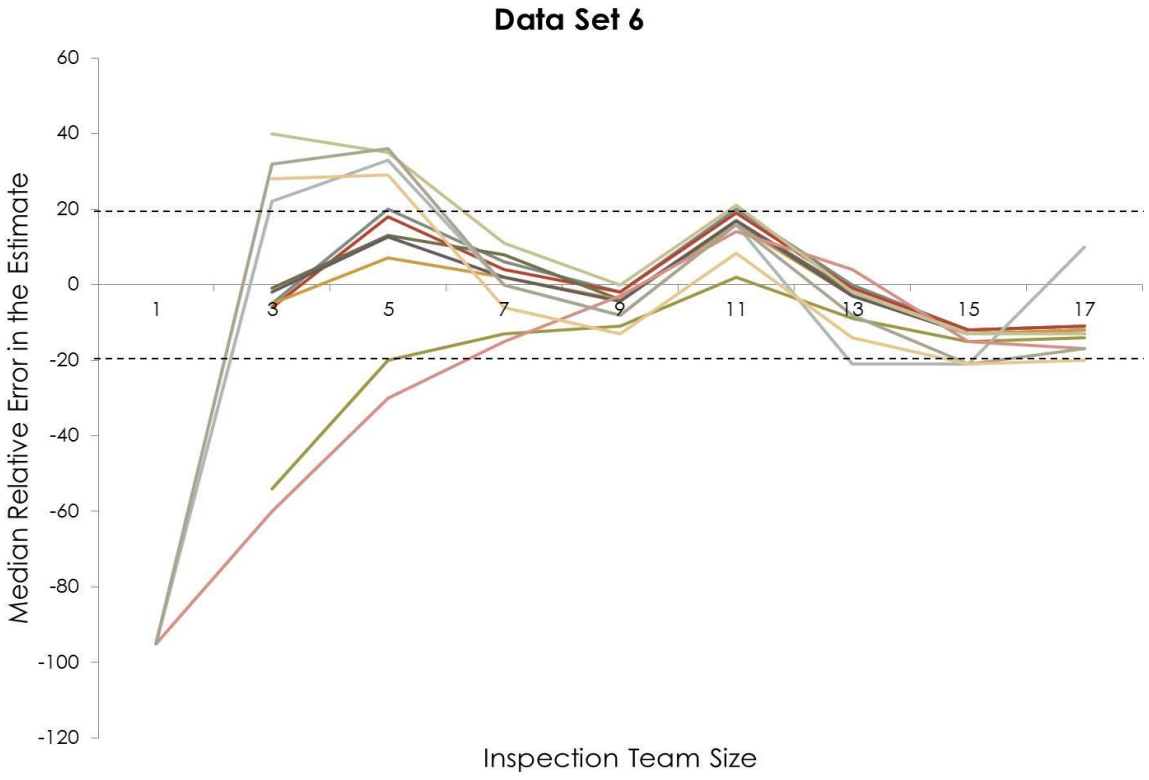

Figure 8 - Median Relative Error in the Estimates vs. Inspection Team Size for Data Set 6

The same analysis process (described in Sections 5.1 and 5.2) is used to evaluate the accuracy of CR estimators on Data set 6. The accuracy (i.e., the median relative error) value for each CR estimator across all team sizes (1-17) is shown in Figure 8. As seen with previous data sets, there is direct improvement in the median estimate with the increase in number of inspectors.

Similar to previous analysis, we analyzed the accuracy (median estimate) and precision (25^{th} largest estimate and 75^{th} largest estimate) to determine the minimum number of inspectors

Table 7 - Number of Inspectors Required for Achieving Different Levels of Estimation while comparing Accuracy + Precision vs. Accuracy for CR Estimators

Accuracy + Precision + Failure Rate

Estimators	0% +/- 10%	+/- 20%	+/- 30%
Mo-CMLE	The CR Estimates obtained Never reached these Relative Error (R.E) Level	16	6
Mo-UMLE		13	6
Mo-EE		16	7
Mt-CMLE		16	6
Mt-UMLE		13	7
Mt-EE		13	7
Mh-SC		14	7
Mh-JK		Never within this range	7
Mh-EE		13	8
Mth-SC		14	7
Mth-EE		17	8

Accuracy

Estimators	-20%
Mo-CMLE	3
Mo-UMLE	3
Mo-EE	6
Mt-CMLE	3
Mt-UMLE	4
Mt-EE	3
Mh-SC	14
Mh-JK	7
Mh-EE	13
Mth-SC	14
Mth-EE	17

required for achieving satisfactory estimate. Table 7 compares the number of inspectors required to achieve varying levels of estimation accuracy and precision (the left side of Table 7) vs. the number of inspectors required to achieve satisfactory (i.e., +/- 20%) estimation accuracy alone (the right side of table 7). The major results from Figure 8 and Table 7 are discussed as follows.

Comparing the results based on the estimate's accuracy and precision (left side) vs. the results obtained from median estimates alone (right side), the M_o and M_t estimators need 3 to 6 inspectors to achieve a median estimate within +/- 20% range, whereas these same CR estimators need 13 to 16 inspectors when combining the median and the interquartile range of the estimates. This result is similar to the results obtained from a larger data set 1 in more than one way:

a) The number of inspectors required to achieve an accurate and precise estimate varies from 13-17 for the CR estimators (except JK estimator) in data set 6 and is very close to the inspector count (that varies from 12-20) obtained from data set 1.

b) The inspector count based only on the median estimate is always less than when calculated using the median estimate along with the range of the middle 50% of the estimates. Furthermore, this difference is less for the CR estimators belonging to the M_h and M_{th} models as compared to the M_o and M_t models. This result is consistently true for data sets 6 and 1.

c) The JK estimator achieved median estimate with a relative error of 20% with only 7 inspectors in data set 6 which is close the inspector count (of 10 inspectors) in data set 1. However, when considering the accuracy and precision values, the JK estimator failed to achieve an estimate within 20% with 17 inspectors and is consistent with the results obtained from data set 1.

d) The EE estimators show failure to produce an estimate for both the data set (1 and 6) as well. However the failure rate is lower for these estimators in data set 6 as compared to the results with previous data sets.

Based on the above discussion, SC estimator for M_h and M_{th} models and the CMLE and UMLE estimators for M_o and M_t models are recommended for use. The recommendation of the SC estimators is again consistent with our earlier result with data set 1.

6. THREATS TO VALIDITY

We faced the following threats to validity in our study.

C**onclusion Validity:** The threat due to the heterogeneity of participants was not controlled across all the data sets. The inspectors in Data Set 1 were Microsoft professionals whereas the inspectors in Data Set 2 through 6 were undergraduate and graduate students.

External Validity: Data sets 2 through 5 were obtained from a course setting where the participants worked with a real client to develop requirements for a system that they later implemented. However, there remains a threat because the participants were all undergraduate students in an educational setting and likely do not represent professional developers. Also, the nature of faults made by students during development can differ from the faults made by software professionals. To mitigate this validity threat, Data Sets 1 and 6 were industrial strength requirement documents that contained realistic defects. In Data Set 1, the realistic defects were seeded into the document rather than being naturally occurring (as in Data Set 6). But, the defects were seeded by researchers who had no knowledge that results would be used for a capture-recapture study. Therefore, the defects were not seeded in such a way to specifically benefit a capture-recapture analysis.

Construct Validity: The actual number of defects present in Data Sets 2 through 6 is not known and might actually be higher than the assumed defect count (i.e., the total number of defects found after two inspections). Therefore, it is possible that teams could have made more errors during development that were not detected during the inspection. Also, we did not collect any data regarding faults that might have occurred during implementation. During the original inspection studies (from which data sets 2 through 6 were analyzed), the CR models were not used, the inspectors' subjective opinion regarding the remaining defects after the second

inspection (which was all that was available) was collected. The inspectors agreed that they had located all the defects present in the artifact during second inspection, ruling out any need of further inspection. So, the inspection process was stopped.

Internal Validity: To reduce the threat of using a small number of inspectors, the number of inspectors used in Data Set 1 is the largest used in any previous study of this type. Also, there could be an effect of the larger number of average faults found by inspectors in some data sets on the estimation performance. However, this is outside the scope of this study and we were only interested in analyzing the effect of the overlap of the faults found by multiple inspectors. Additionally, to control the variability of the inspection techniques, we used the data from first inspection for all the six data sets in which all the inspectors used the same inspection technique (i.e., fault checklist) to detect defects.

7. DISCUSSION OF RESULTS

This section brings the results from all the six data sets in light of the original research goals. This section discusses the major finding and recommendation about the minimum number of inspectors required for the CR models and estimators to achieve reliable estimates. The major findings from this study are also compared with the earlier findings from software engineering and biology.

7.1 Summary of Findings and Recommendation

This study evaluated the effect of the number of inspectors on the capture-recapture models and estimators based on the accuracy, precision, and failure rate of their estimators. Based on the results from all the six data sets, we present a summary of major findings and recommend the best estimator(s) as follows:

Effect of Inspection Team Size: Across all the data sets, an increase in the inspection team size improves the accuracy and precision of all the CR estimators. Also, the failure rates of CR estimators improve with increasing inspection team size. However, there are certain CR estimators (i.e., EE estimators for M_o and M_{th} models) that failed to produce an estimate with increase in inspection team size (even when they had produced an estimate with less number of inspectors). Therefore, we do not recommend the EE estimators for use.

Accuracy vs. Accuracy + Precision: The minimum number of inspectors required to achieve a satisfactory median estimate is very different from the minimum number of inspectors required to achieve an accurate as well as precise estimate. This was true across all the data sets and is an interesting result due to the following reasons:

a) It highlights an aspect of previous research in software engineering, where findings are based only on the median estimates.

b) When using the accuracy values on data sets, in one of the data set (i.e., Data Set 6), our findings are similar to some of the previous findings, that six inspectors are enough to produce an estimate with a relative error of +/- 20%. However, when using both the accuracy and precision values, we find that the CR estimators need a minimum of 13 or 17 inspectors to produce an estimate within +/- 20%. Furthermore, with larger data sets (e.g., with 73 inspectors) used in our studies, the difference in the *accuracy* vs. *accuracy + precision* values at +/- 20% is not as wide as with smaller data sets.

c) Another result is that, with smaller data sets (with six or eight inspectors) as used in the previous research, it is hard to objectively find the minimum number of inspectors required to achieve an accurate and precise estimate. This is because (as seen in data sets with larger inspectors), the point of six and eight inspectors represent an area of huge variability where the estimate tend to vary from a high negative r.e. to a positive r.e with an increment of just one inspector. The results in b) and c) show that to properly evaluate the effect of the inspection team size on the CR estimators; we need data sets with large number of inspectors.

Best Estimator(s): Contrary to previous findings, we find that M_h-JK is not the best estimator. The jackknife estimator needs an extremely large number of inspectors (i.e., 46 as in Data set 1) for it to produce an accurate and precise estimate within +/- 20%. Unlike other estimators, the Jackknife estimator shows an inconsistent improvement in the precision with increasing inspection team size. Out of all the CR estimators, we recommend the SC estimators for M_h and M_{th} models to use as they need the least number of inspectors (12 to 14) to achieve an

accurate and precise estimate within +/- 20%. The UMLE and CMLE estimators were second best as they need 13 to 20 inspectors to achieve a satisfactory estimate. While the EE estimators for M_h model performed well, we do not recommend it because of its failure to produce an estimate for some of the virtual inspection data sets.

7.2 Relevance to Software Organizations

Software organization needs to decide whether or not to re-inspect an artifact based on an estimate of the number of defects remaining after an inspection. To accurately use capture-recapture models, it is imperative for them to know the relative performance of the different estimators and to select a small number to use on their data sets that provide reliable estimates. Since the estimates improve with more inspectors, information about the minimum number of inspectors required to achieve satisfactory estimates is with varying levels of accuracy and precision for different estimators can help project managers to plan and manage the inspection process relevant to organizations. The information regarding the minimum number of inspectors required for achieving estimates with varying levels of accuracy and precision can help project managers to better plan and manage the inspection process.

7.3 Comparison with Previous Findings in Biology and Software Engineering

Table 8 compares the findings from this study with the previous findings from the application of capture-recapture models in software engineering and biology and wildlife research. Findings from this study confirm some of the previous findings but provide additional insights into the performance of estimators.

Table 8 - Comparison of Finding

S.No	Our Study	Software Engineering	Biology and Wildlife
1	All models underestimate or overestimate with a small number of inspectors and estimation improve s with more inspectors	All models generally underestimate but estimates improve with more inspectors	All models generally underestimate but estimates improve with more trapping occasions
2	M h -JK with inspectors <= 46 show severe variation. Contrary to earlier findings, M h -JK is highly unreliable with <=46 inspectors	For 4 or more inspectors, Mh-JK is recommended as the best estimators for most of the studies in software engineering research	Mh-JK severely overestimates in case of few trappings, but provide good estimates if the overlap of animals caught at different trappings is large
3	The minimum number of inspectors required to obtain stable estimates within relative error of +/- 20% can vary from 13 to 20 depending on the estimator	The minimum number of inspectors required for getting accurate estimates is four (4), however no standard measurement of accuracy is provided	5 trappings occasions are recommended as minimum number, but 7 or 10 can provide better estimates, however no justification of this recommendation is provided
5	The UMLE estimators are better than the JK estimator. The UMLE estimator underestimates with fewer number of inspectors as compared to JK estimator but is more precise	An initial study shows that UMLE for Mt (underestimates) is better than Mh -JK that overestimates. On the contrary. another finding show that Mo-MLE overestimates	MLE estimators for Mo and Mt type models produce highly inaccurate estimates as compared to M h models
6	The failure rate is high if the number of inspectors is less than four or five. The EE estimators had failure rates for larger inspectors too	Failure rate is high for number of inspectors less than 4	Failure rate is high for few inspectors
7	Sample coverage for Mh and Mth type are the best estimators		Sample coverage estimator performs less well than MLE estimator

8. CONCLUSION AND FUTURE WORK

Based on the results provided in this paper, the capture-recapture models can help manage the quality of software artifacts. Software organizations can use the results in this paper about the number of inspectors required for achieving defect estimates at varying levels of estimation accuracy and precision as needed.

Software project managers also need to make a tradeoff between the costs involved in using more inspectors for them to be able to use the inspection data to make objective post-inspection decisions. In addition, if they decide on re-inspections, the cost effectiveness of doing a re-inspection should be examined with respect to the cost vs. the benefits of finding the defects estimated to be remaining. We have recently started working on providing guidance on how to appropriately use the cost metrics to evaluate the cost-effectiveness of software inspections and post-inspection decisions based on the CR estimates

9. REFERENCES

[1] Ackerman, A., Buchwald, L., and Lewski, F., "Software Inspections: An Effective Verification Process." IEEE Software, 1989. 6(3): 31-36.

[2] Briand, L.C., Emam, K.E., and B.G.Freimut. "A Comparison and Integration of Capture-Recapture ". In Proceedings of the 9th International Symposium on Software Reliability Engineering. 1998. Paderborn, Germany: 32-41.

[3] Briand, L.C., Emam, K.E., Freimut, B.G., and Laitenberger, O., "A Comprehensive Evaluation of Capture Recapture Models for Estimating Software Defect Content." IEEE Transactions on Software Engineering, 2000. 26(6): 518-539.

[4] Burnham, K.P. and Overtom, W.S., "Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals." Biometrica, 1978. 65: 625-633.

[5] Chao, A., "Estimation the population Size for Capture-Recapture Data with Unequal Catchability." Biometrics, 1987. 43(4): 783-791.

[6] Chao, A., "Estimating Animal Abundance with Capture Frequency Data." Journal of Wildlife Management, 1988. 52(2): 295-300.

[7] Chao, A. and Yeng, H.C., Program CARE-2 (for Capture-Recapture Part.2), http://chao.stat.nthu.edu.tw, 2003.

[8] Darroch, J.N., "The Multiple-Recapture Conensus 1: Estimation of a Closed Population." Biometrika, 1958. 45: 343-359.

[9] Ebrahimi, N.B., "On the Statisctical Analysis of the Number of Errors Remaining in a Software Design Document after Inspection." IEEE Transactions on Software Engineering, 1997. 23(8): 529-532.

[10] Eick, S., Loader, C., Long, M., Votta, L., and Weil, S.V. "Estimating Software Fault Content Before Coding". In Proceedings of the 14th International Conference on Software Engineering. 1992. Melbourne, Australia: ACM Press: 59-65

[11] Eick, S., Loader, C., Weil, S.V., and Votta, L. "How Many Errors Remain in a Software Design after Inspection". In Proceedings of the 25th Symposium on the Interface. 1993:

[12] El-Emam, K., Laitenberger, O., and Harbrich, T., "The Application of Subjective Estimates of Effectiveness to Controlling Software Inspections " Journal of Systems and Software, 2000. 54(2): 119-136.

[13] El-Emam, K. and Laitenberger, O., "Evaluating Capture-Recapture Models with Two Inspectors." IEEE Transactions on Software Engineering, 2001. 27(9): 851-864.

[14] Lee, S.M. and Chao, A., "Estimating Population Size via Sample Coverage for Closed Capture-Recapture Models." Biometrics, 1994. 50: 88-97.

[15] Miller, J., “Estimating the Number of Remaining Defects after Inspection.” Software Testing, Verification and Reliability, 1999. 9(3): 167-189.

[16] Otis, D., Burnham, K., White, G., and Anderson, D., “Statistical Inference from Capture Data on Closed Animal Population.” Wildlife Monograph, 1978. 64: 1-135.

[17] Petersson, H., Thelin, T., Runeson, P., and Wohlin, C., “Capture-Recapture in Software Inspections after 10 Years Research - Theory, Evaluation and Application.” Journal of Systems and Software, 2003.

[18] Runeson, P. and Wohlin, C., “An Experimental Evaluation of an Experience-Based Capture-Recapture Method in Software Code Inspections.” Empirical Software Engineering: An International Journal, 1998. 3(4): 381-406.

[19] Shull, F., Carver, J., and Travassos, G. “An Empirical Methodology for Introducing Software Processes”. In Proceedings of Joint 8th European Software Engineering Conference and 9th ACM SIGSOFT Foundations of Software Engineering. 2001. Vienna, Austria: 288-296.

[20] Thelin, T. and Runeson, P. “Capture-Recapture Estimators for perspective Based Reading - A Simulating Experiment”. In Proceedings of the International Conference on Product Focused Software Process Improvement. 1999

[21] Thelin, T., Petersson, P., and Runeson, P., “Confidence Intervals for Capture-Recapture Estimations in Software Inspections.” Journal of Information and Software Technology, 2002. 44(12): 683-702.

[22] Weil, S.V. and Votta, L., “Assessing Software Designs Using Capture-Recapture Methoda.” IEEE Transactions on Software Engineering, 1993. 19(11): 1045-1054.

[23] White, G.C., Anderson, D.R., Burnham, K.p., and Otis, D.l., Capture-Recapture and Removal Methods for Sampling Closed Populations, Los Alomos National Laboratory, 1982.

[24] Wohlin, C., Runeson, P., and Brantestam, J., “An Experimental Evaluation of Capture-Recapture in Software Inspections.” Software Testing, Verification and Reliability, 1995. 5(4): 213-232.

[25] Wohlin, C. and Runeson, P. “Defect Content Estimation from Review Data”. In Proceedings of the 20th International Conference on Software Engineering. 1998. Kyoto, Japan: IEEE Computer Society Press: 400-409

[26] Yip, P.S.F., “A Martingale Estimating Equation for a Capture-Recapture Experiment in Discrete Time.” Biometrics, 1991. 47: 1081-1088.

[27] **Walia, G.**, Carver, J. and Nagappan, N. “The Effect of the Number of Inspectors on the Defect Estimates Produced by Capture-Recapture Models.” *Proceedings of the 30th International Conference on Software Engineering - ICSE'2008*. May 10-18, 2008. Leipzig, Germany. p. 331-340.

[28] **Walia, G.**, Carver, J. and Philip, T. "Requirement Error Abstraction and Classification: A Control Group Replicated Study." *Proceedings of the 18th IEEE International Symposium on Software Reliability Engineering - ISSRE'2007*. November 5-9, 2007. Trollhättan, Sweden. pp. 71-80.

[29] **Walia, G**., and Carver, J. "Requirements Error Abstraction and Classification: An Empirical Study." *Proceedings of the 2006 International Symposium on Empirical Software Engineering - ISESE'2006*. Sept. 21-22, 2006. Rio de Janeiro, Brazil. pp. 336-345.