Assessing Software Designs using Capture-Recapture Methods

Scott A. Vander Wiel
AT&T Bell Laboratories, Murray Hill

Lawrence G. Votta
AT&T Bell Laboratories, Indian Hill

June 5, 1993

Abstract

Sometimes complex software systems fail because of faults introduced in the design stage of the development process. Design reviews can remove some of these faults but often a few remain undiscovered until the software is developed further. The number of faults not discovered by the design review can be estimated by using capture-recapture methods. Since these methods were developed for wildlife population estimation, the assumptions used to derive them don't match design review applications. We report on a Monte-Carlo simulation to study the effects of broken assumptions on maximum likelihood estimators (MLEs) and jackknife estimators (JEs) of faults remaining. We find that the MLE performs satisfactorily if faults are classified into a small number of homogeneous groups. Without grouping, the MLE can perform poorly but it generally does better than the JE.

Keywords: software quality assurance, design faults, design reviews, statistical estimation, Monte-Carlo simulation

1 Software Designs and Capture-Recapture Models

Many complex systems are controlled by computer software. Some examples are the systems for placing long distance telephone calls, ticketing airline reservations, and verifying consumer credit. These systems typically function so well that we are unaware of their complexities. Sometimes, however, they fail because of software faults. Faults are often introduced into systems at the design stage of the development process [5]. Unfortunately, design reviews usually do not remove all of them. It is important to eliminate these design faults as quickly as possible because they become much more expensive to fix as a software system proceeds through development (see [1], pp. 39–41).

To better estimate the number of remaining faults after design review, we have integrated capture-recapture[1] methods into the design process in AT&T's International Switching Division. Before a peer review meeting for the design of a software module, a design document is distributed to several reviewers who read it and note "issues" that they think should be resolved before the software feature is developed further. We refer to these issues as faults. At the review meeting data is collected showing which reviewers discovered which faults. This data is then used to estimate the number of faults that remain undiscovered. These estimates can then be used in managing the quality at the design stage of the software development process.

Estimating the number of *remaining* faults is new to the design process at AT&T which, at one time, specified both upper and lower limits on the number of faults *found* per page during design review meetings in order for a design to be acceptable. The intent was to ensure that design reviews were thorough (hence the lower limit) and that design documents were of high quality (the upper limit). This practice, however, encouraged reviewers to find a "passing" number of faults regardless of the quality of the design document [3].

The ultimate goal is to prevent design faults altogether. As a first step we concentrate on developing methods to detect designs of questionable quality earlier in the process by estimating the number of faults that go undetected after a design review. This is valuable for deciding whether the design is of high enough quality for the software to be developed further.

Our approach treats the faults found by reviewers preparing for a design review as data from a capture-recapture sampling scheme. We use a Monte Carlo simulation to investigate the inaccuracies of the capture-recapture estimators due to assumption violations when faults have varying detection probabilities and reviewers have different capture probabilities. Although we would like to use data from real world design reviews to perform this study, it is impossible. We can not control the fault detection and reviewer capture probabilities in design reviews, nor, could we ever hope to obtain the number of reviews required to perform a statistically significant study.

Our study shows that maximum likelihood and jackknife estimators are not robust to assumption violations, but one can improve the maximum likelihood estimator by analyzing different types of faults separately. Further work should be done to determine a reliable classification method.

[1] Capture-recapture methods were developed for estimating the sizes of wildlife populations. Good accounts of the theory and practice of these methods are given in [2] and [8] respectively.

1.1 Background

Eick et al. [3] describe some preliminary results from using capture-recapture estimators in AT&T's International Switching Division. They present various alternative capture-recapture estimators and show how to test some of the model assumptions.

1.2 Overview of the paper

This paper defines a maximum likelihood estimator (MLE) and a jackknife estimator (JE) of the number of undiscovered faults f_0 and compares their behavior in situations where the usual model assumptions are violated. We chose to compare these estimators because they are developed from two very different sets of assumptions neither of which hold in practice. We wanted to know whether one estimator performed better than the other under assumptions more realistic to software development.

The main result is that the MLE can severely misestimate the remaining number of faults unless, they are classified into groups. Grouping does not improve the JE. We also recommend setting upper confidence bounds on the number of remaining faults using a likelihood ratio method.

This paper is organized as follows. Section 2 introduces the maximum likelihood and jackknife estimators and discusses the assumptions used to derive them. Section 3 describes the simulation and discusses the behavior of the estimators when fault type is ignored as well as when faults are grouped by type. Section 4 is a more thorough study of the simulation results. Section 5 summarizes and concludes.

2 Estimators of the Number of Undiscovered Faults

We compare 2 estimators used by biologists for capture-recapture studies of wildlife populations; but we use them to estimate software design faults. The MLE is derived from assumptions that accommodate different reviewer detection probabilities for observing a fault, but it treats all faults as probabilistically identical. The JE is exactly the opposite. It is intended for situations in which the reviewers are identical but the faults may have different detection probabilities. Both techniques have their advantages and disadvantages, as we will investigate in Section 3. For further details and alternative estimators see [7].

We use the following problem description and notation in the rest of the paper. There is an unknown number, N, of faults in a design document. In preparation for a review meeting by t reviewers, $n \leq N$ of the faults are discovered—some by more than 1 reviewer. Let n_j denote the number discovered by reviewer j ($j = 1, \ldots, t$) and let f_k denote the number discovered by exactly k reviewers ($k = 0, \ldots, t$). Then f_0 is the number of undiscovered faults, $n = f_1 + \cdots + f_t$, and $N = f_0 + f_1 + \cdots + f_t$.

2.1 Maximum likelihood

The MLE used in this paper is based on a probability model in which the events that reviewer j discovers fault i ($i = 1, \ldots, N$; $j = 1, \ldots, t$) are independent with probabilities $p_{ij} = p_j$ that depend only on the reviewer. We describe this by saying that all faults are probabilistically independent and identical. The reviewers act independently but different reviewers may have different probabilities for discovering faults.

Under this model the log-likelihood maximized over $p_1, \ldots, p_t$ as a function of f_0 is

$$\begin{aligned} \mathcal{L}(f_0) &= \ln \binom{N}{n} + \sum_{j=1}^{t} n_j \ln n_j \\ &\quad + \sum_{j=1}^{t} (N - n_j) \ln(N - n_j) - Nt \ln N \end{aligned}$$

where $N = f_0 + n$. The maximizer $\hat{f}_0$ of $\mathcal{L}(f_0)$ is the MLE. See [4] for a derivation and further details. Technically $\mathcal{L}$ should be defined only for nonnegative integers f_0. Standard practice, however, is to replace factorials with gamma functions, do calculations as if f_0 could be any nonnegative real and round results to integers when reporting final estimates and confidence bounds.

2.2 Jackknife

Burnham and Overton [2] developed an estimator of population size based on the generalized jackknife. Their jackknife estimator (JE) is intended to do well when faults have varying difficulty. Their derivation uses a probability model in which the difficulty of each fault is described by its discovery probability p_i, $i = 1, \ldots, N$. The p_i's are assumed to be a random sample from an unknown distribution. Given the p_i's the events that reviewer j discovers fault i are independent with probabilities $p_{ij} = p_i$ depending only on the fault index i. We say that reviewers are probabilistically independent and identical. Faults have differing difficulties (p_i's) but the difficulty of any given fault is the same for all reviewers.

The JE derived from this model has the form

$$\hat{f}_0 = \max\{0, a_1 f_1 + \cdots + a_k f_k\}$$

where $k \leq t$ is the jackknife order and $a_1, \ldots, a_k$ are given constants (depending on t and k). Full details as well as a method for selecting k based on the data are givin in [2]. The simulation results discussed below incorporate their method for selecting k.

2.3 Grouping faults

If there are several types of faults with different discovery probabilities, the MLE can perform poorly because it treats all faults the same—we have *pooled* them together to estimate f_0. In this case it is helpful to classify the faults into a small number of groups and estimate faults separately within each group. These groups should be formed so the for a given reviewer and a given group, all the issues should have identical capture probabilities. Our simulation has two types of faults, A and B. We study MLE and JE estimators formed by estimating the two populations of faults separately and then adding the results. If $\hat{f}_{0\mathrm{A}}$ and $\hat{f}_{0\mathrm{B}}$ are MLEs of undiscovered type A and type B faults respectively, then the *grouped* MLE is

$$\hat{f}_0 = \hat{f}_{0\mathrm{A}} + \hat{f}_{0\mathrm{B}}.$$

We also study the grouped JE which is defined similarly.

2.4 Confidence bounds

For estimators with small bias it is interesting to compare coverage probabilities for confidence bounds constructed using various approximations. We compare 3 methods for constructing upper confidence bounds on f_0 and we recommend one of them. An upper confidence bound is the largest value of f_0 that the data can credibly support. An upper bound that is far greater than $\hat{f}_0$ indicates that the data support a wide range of possible values for f_0. Though lower bounds are just as easy to calculate, upper bounds are more useful to a software developer who wants to be confident that no more than a few faults remain undiscovered in a design.

Wald confidence bounds for f_0 are based on the asymptotic normal distribution of the MLE. An approximate upper $100(1-\gamma)\%$ confidence bound is

$$\hat{f}_0 + z_\gamma s_\mathrm{W}$$

where $\hat{f}_0$ is the MLE, z_γ is the $1-\gamma$ quantile of the standard normal distribution, and

$$s_\mathrm{W} = (n+\hat{f}_0)\left[\prod_j \hat{q}_j^{-1} - \sum_j \hat{q}_j^{-1} + t - 1\right]^{-1}$$

with $\hat{q}_j = (n+\hat{f}_0 - n_j)/(n+\hat{f}_0)$ [2].

Likelihood ratio confidence bounds are based on the asymptotic distribution of the likelihood ratio statistic. An approximate upper $100(1-\gamma)\%$ confidence bound is

$$\sup\{f_0 : \mathcal{L}(\hat{f}_0) - \mathcal{L}(f_0) < \chi^2_{1,\gamma}/2\}$$

where $\hat{f}_0$ is the MLE and $\chi^2_{1,\gamma}$ is the $1-\gamma$ quantile of the χ^2_1 distribution. In instances where the distribution of the MLE is well centered over that of f_0 it is interesting to compare coverage probabilities for Wald and Likelihood Ratio type confidence bounds.

An approximate upper $100(1-\gamma)\%$ confidence bound [2] for f_0 as

$$\hat{f}_0 + z_\gamma s_\mathrm{J}$$

where $\hat{f}_0$ is the JE and

$$s_\mathrm{J} = \frac{n}{n-1}\left[\sum_{j=1}^k a_j^2 f_j - \frac{1}{n}\left(\sum_{j=1}^k a_j f_j\right)^2\right].$$

In instances where both the MLE and the JE have small bias it is interesting to compare coverage probabilities for Wald, Likelihood Ratio and Jackknife confidence bounds.

When faults are grouped according to type, Wald confidence bounds for f_0 can be computed as

$$(\hat{f}_{0\mathrm{A}} + \hat{f}_{0\mathrm{B}}) + z_\gamma\sqrt{s^2_{\mathrm{WA}} + s^2_{\mathrm{WB}}}$$

where our notation is an obvious extension of that above.

Likelihood Ratio confidence bounds are computed from

$$\sup\{f_0 : \mathcal{L}_\mathrm{A}(\hat{f}_{0\mathrm{A}}) + \mathcal{L}_\mathrm{B}(\hat{f}_{0\mathrm{B}}) - \max_{f_{0\mathrm{A}}}(\mathcal{L}_\mathrm{A}(f_{0\mathrm{A}}) + \mathcal{L}_\mathrm{B}(f_0 - f_{0\mathrm{A}})) < \chi^2_{1,\gamma}/2\}.$$

Jackknife confidence bounds are given by

$$(\hat{f}_{0\mathrm{A}} + \hat{f}_{0\mathrm{B}}) + z_\gamma\sqrt{s^2_{\mathrm{JA}} + s^2_{\mathrm{JB}}}.$$

3 Simulation

In this section, we describe a simulation study comparing the behavior of the MLE and JE. Otis et al. [7] give extensive simulation results on various estimators of population size under many probability models. They include results for the MLE and JE. Their focus, however, is on estimating the *total* population size whereas we focus on estimating the size of the *undiscovered* population. Thus, we compare $\hat{f}_0$ to the true value f_0 while Otis et al.compare $\hat{N} = \hat{f}_0 + n$ to the true value N. This difference makes our results seem more alarming—estimating only 5 remaining faults when there are actually 10 sounds worse than estimating 95 total faults when there were 100. Two other substantive differences between our simulation and that of Otis et al. are as follows. Our study points (see Subsection 3.1) represent document reviews while theirs represent capture-recapture wildlife surveys. The sets of fault discovery probabilities we judged to be typical for design reviews are not included in the sets of capture probabilities Otis et al. studied. Secondly, we study 3 methods for setting confidence bounds. Two of these, including the one we recommend, were not studied by Otis et al.

We describe only 3 of 25 study points in this section. The 3 examples, however, show the two main results of the paper: the MLE and JE can severely misestimate the number of remaining faults in mixed fault populations, and this misestimation can be largely corrected for MLE by categorizing the faults and estimating from each group separately (Subsection 3.2). Section 4 presents results for all the simulated study points.

3.1 Cases studied

Often in our observations and follow-up interviews of requirements and design reviews (we have observed over 50 reviews and have interviewed over 100 reviewers), we observed that

1. many reviewers think that faults occur in at least two categories; easy to observe (not requiring much thought) and hard to observe (requiring much thought);

2. one or two reviewers contribute a large percentage of the faults observed; and

3. many reviews have a specialist—one reviewer whose sole goal is to identify one type of fault.

We simulated four different cases[2] as shown in Table 1. They are representative of design and requirement document reviews. The last column of the table references each case to one of the three observations listed above. The four cases allow us to probe the biases that the MLE and JE may be susceptible to when applied to software design reviews.

For all cases, we assumed 5 reviewers for each review with 100 faults to be discovered. Faults were of two types: N_{A} of Type A and N_{B} of Type B. The mix of faults (N_{A},N_{B}) varied over the set $\{(100,0),(75,25),(50,50),(25,75),(0,100)\}$. For each ($N_{\mathrm{A}}$,$N_{\mathrm{B}}$) pair the vectors $\mathbf{p}_{\mathrm{A}}$ and $\mathbf{p}_{\mathrm{B}}$ of discovery probabilities for the 5 reviewers varied over the 4 cases shown in Table 1.

Cases 1 and 2 model when the reviewers on a team are almost identical in their ability to find faults but some faults are more difficult to find than others. In Case 1 each reviewer has a small (10%) chance of finding Type A faults and a medium (60%) chance of finding Type B faults. In Case 2 they have medium (40%) and high (90%) chances respectively. Columns 4 and 5 of Table 1 list the expected percent of undiscovered faults of each type. This ranges from near 60% when faults are difficult to find (Case 1, Type A) to near 0 when faults are easy to find (Case 2, Type B.)

Cases 3 and 4 model differences among reviewers in addition to the differences between fault types. In Case 3 two reviewers specialize in finding Type A faults and two others specialize in finding Type B faults. Just over 2% of faults of either type remain undiscovered on average. In Case 4 Type A faults are difficult for all reviewers to find and about 22% go undiscovered on average. Type B faults are similar except that one reviewer can find most of them (80% on average).

Crossing the 4 cases in Table 1 with the 5 (N_{A},N_{B}) pairs gives 20 study points. For each study point 1000 reviews were simulated. The study points with $N_{\mathrm{A}} = 0$ or $N_{\mathrm{B}} = 0$ satisfy the assumptions used to derive the MLE. The other study points do not conform to the assumptions for either the MLE or the JE. The study points corresponding to Cases 1 and 2 are, however, similar to those used to derive the JE.

For a given Case and (N_{A}, N_{B}) pair we generated data sets representing 1000 reviews. Each dataset consisted of two matrices (one $N_{\mathrm{A}} \times 5$ and one $N_{\mathrm{B}} \times 5$) of independent Bernoulli variates indicating which reviewers discovered which faults. From each matrix we calculate the summary statistics needed to form the various estimates and confidence bounds for faults remaining. These summary statistics are n, $(n_1,\ldots,n_5)$ and $(f_1,\ldots,f_5)$. The summary data were then used to calculate separate estimates of f_0 for type A and type B faults and a third estimate based on the pooled data for both fault types. This gave 6 estimates—3 for the MLE and 3 for the JE. Also for each data set we computed 9 confidence bounds: Wald, Likelihood ratio and Jackknife bounds for type A, type B and pooled data. The next section summarizes our findings.

[2] We use Case to refer to a choice of fault reviewer detection probabilities vectors ($\mathbf{p}_{\mathrm{A}}, \mathbf{p}_{\mathrm{B}}$). The term study point refers to a choice of the numbers of each type of fault (N_{A},N_{B}) as well as the detection probability vectors ($\mathbf{p}_{\mathrm{A}}, \mathbf{p}_{\mathrm{B}}$).

3.2 Behavior of MLE and JE

Figures 1, 3 and 4 show scatter plots of the estimated number of undiscovered faults $\hat{f}_0$ plotted against the actual number f_0. Figure 1 displays both the MLE and the JE for a study point with only one fault type. The data sets for Figures 3 and 4 had two types of faults. This was ignored in Figure 3 but not in Figure 4. We discuss the plots in more detail below, but notice that both estimators in Figure 3 have significant biases. Comparing the left panels of Figure 3 and Figure 4 shows how much this bias can be reduced when MLEs are calculated separately for each fault type.

All of our scatter plots have $\log(1+x)$ scales for both $\hat{f}_0$ and f_0. This makes, for example, a 10% increase from $\hat{f}_0 = 10$ to $\hat{f}_0 = 11$ roughly the same vertical distance as a 10% increase from $\hat{f}_0 = 100$ to $\hat{f}_0 = 110$. (A $\log(x)$ scale would make the comparison exact but then $\hat{f}_0 = 0$ could not be included.) The plotted points have been jittered for clarity but the vertical bands of points in the plots remind us that f_0 is an integer. In the comparisons below we use the median estimation error $\mathrm{med}(\hat{f}_0 - f_0)$ to summarize the typical amount by which $\hat{f}_0$ over- or under-estimates f_0. The reason we use the median rather than the mean is that the mean estimation error of the MLE is theoretically infinite because there is always a small probability that $\hat{f}_0$ will be infinite. We discuss this further in subsection 4.2.

3.2.1 Study point with 1 fault type

Figure 1 shows plots of $\hat{f}_0$ vs. f_0 for Case 4 with $(N_{\mathrm{A}}, N_{\mathrm{B}}) = (0,100)$. The MLE is plotted in the left panel and the JE in the right one. Only Type B faults are present so this study point obeys the assumptions used to derive the MLE. One reviewer finds 80% of the faults on average. Each of the other four finds an average of 20%.

In this case the MLE behaves much better than the JE. More than half of the MLEs are between 4 below f_0 and 2 above f_0 with a median of 1.1 below. (See Tables 2 and 3 and Figure 5 in Section 4.)

The JE, on the other hand, does poorly. Just over 8 faults go undetected on average yet half of the JEs are from 23 to 39 faults too high with a median of about 30. Thus, the JE often overestimates by a factor of 3 to 4.

Another striking observation from Figure 1 is that neither the MLE nor the JE appears to be correlated with f_0. This

Table 1: Simulation cases

Case	discovery probabilities $\mathbf{p}_{\mathrm{A}}$	$\mathbf{p}_{\mathrm{B}}$	percent undiscovered Type A	Type B	Observation
1	(.1, .1, .1, .1, .1)	(.6, .6, .6, .6, .6)	59.05%	1.02%	1
2	(.4, .4, .4, .4, .4)	(.9, .9, .9, .9, .9)	7.78	0.001	1
3	(.8, .8, .3, .1, .1)	(.1, .1, .3, .8, .8)	2.27	2.27	1,2,3
4	(.1, .3, .3, .3, .3)	(.8, .2, .2, .2, .2)	21.61	8.19	1,2,3

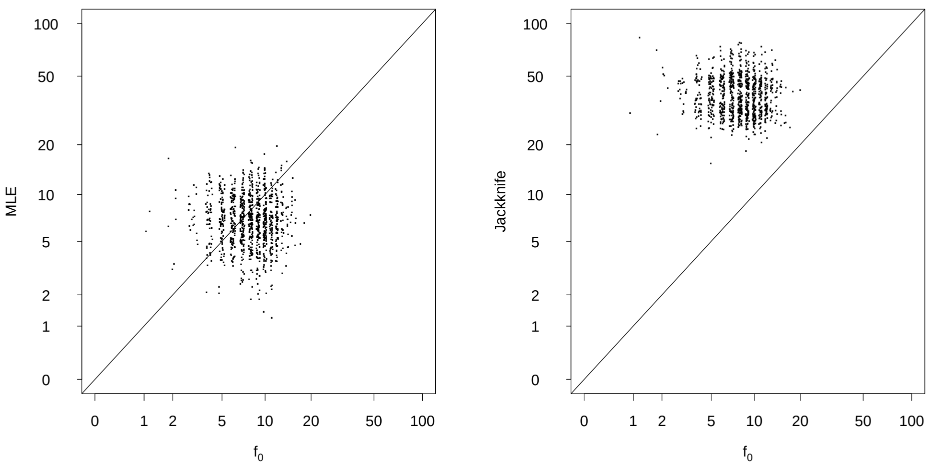

Figure 1: **Simulation Results – One Fault Type.** Plots of $\hat{f}_0$ (MLE on left, jackknife on right) versus f_0 when $N_{\mathrm{B}} = 100$ and $\mathbf{p}_{\mathrm{B}} = (.8, .2, .2, .2, .2)$. The axes are on a $\log(1+x)$ scale. The diagonal line is $\hat{f}_0 = f_0$. For this study point $\mathrm{med}(\hat{f}_0 - f_0)$ is -1.1 faults for the MLE and 29.8 faults for the JE. Neither estimator is well correlated with f_0.

indicates a certain limitation on the information that the observed data contain about f_0. Although the distribution of the MLE is centered over the distribution of f_0, the two quantities are not substantially correlated *when the parameters of the reviews do not change.* In practice, however, each review is different. Some designs have few faults, others have many. The sets of reviewer detection probabilities are different and the number of reviewers is not constant. Figure 2 shows a scatter plot of $\hat{f}_0$ vs. f_0 for the MLE over our entire simulation. When the review conditions change, large values of f_0 tend to produce large estimates $\hat{f}_0$.

3.2.2 Study points with 2 fault types

Figures 3 and 4 are plots of $\hat{f}_0$ for Case 2 with $(N_{\mathrm{A}}, N_{\mathrm{B}}) = (75, 25)$. In Figure 3 neither the MLE nor the JE does well as the centers of the clouds of points are far from the 45 degree line. Here we have treated all faults the same—we have *pooled* them together to estimate f_0.

A potential solution to the population having two types of faults, A and B, is to estimate the populations of faults separately, and then add the results. We call this *grouping* (versus pooling) the faults. In Figure 4 the MLE is significantly improved as the center of the cloud is now very close to the 45 degree line. The grouping of faults has virtually no effect on the JE. The right hand plots of Figures 3 and 4 are almost identical.

We should note that the improvement comes with a price. First, we have assumed that we can define fault types and correctly classify faults into the types. Miller [6] argues that this can be done accurately for about 7 categories. Furthermore, though dividing faults into more homogeneous groups can greatly improve bias, it will also increase the variability of the estimators. It is important to use fault categories broad enough so that several faults within each category are expected to be found by more that one reviewer. Otherwise the variability of the MLE will be large. (In fact if there is a fault type in which no fault is found by at least 2 reviewers, the MLE for that fault type will be infinite.)

3.3 Confidence bounds

Our study has shown that the median estimation error of the MLE is small if there is only one fault type or if faults are grouped into a small number of types. Having obtained an estimator with low bias it is important to compare different methods for setting upper confidence bounds on f_0.

Return to the study point discussed in subsection 3.2.1 in which $(N_{\mathrm{A}}, N_{\mathrm{B}}) = (0, 100)$, $\mathbf{p}_{\mathrm{B}} = (.8, .2, .2, .2, .2)$. We computed Wald and LR upper 95% confidence bounds for each of

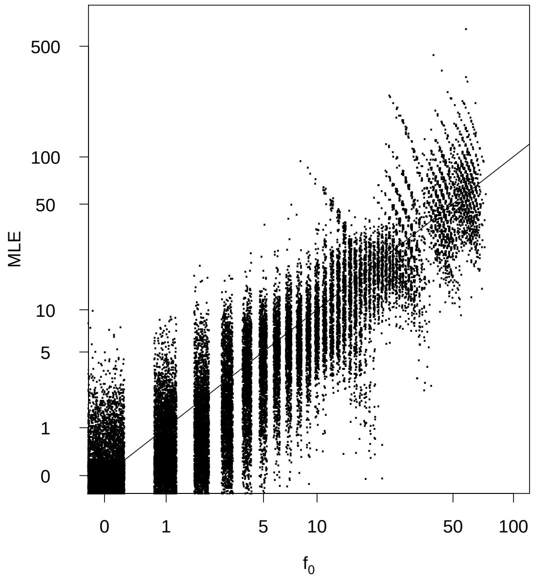

Figure 2: **Simulation Results – All MLE Single Fault Trials.** Scatter plot of the MLE versus f_0 over all simulated data sets with only one type of fault. Large values of f_0 tend to produce large MLEs demonstrating the correlation between f_0 and $\hat{f}_0$

the 1000 simulated data sets. Ideally, f_0 will fall below the upper bounds in 95% of the simulations. The actual percentages are 86.3% for Wald bounds and 99.2% for LR bounds. We refer to these actual percentages as *coverages* and say that an upper bound *fails* when f_0 exceeds the bound. The coverages we report for this study point are typical of the other study points and lead us to the generalization that Wald bounds tend to be anticonservative (they fail too frequently) while LR bounds tend to be conservative (they fail less often than anticipated.)

Using confidence bounds is important because they indicate how much information the data provide to estimate f_0. We recommend using upper LR bounds with the understanding that they tend to fail less often than anticipated and are thus the conservative choice.

4 Further Discussion of Simulation Results

Figure 5 shows box plots of estimation errors from all 20 study points we simulated. The left half of the figure shows results for the MLE and the right half shows results for the JE. Each row of plots corresponds to a different set of discovery probabilities $(\mathbf{p}_{\mathrm{A}}, \mathbf{p}_{\mathrm{B}})$. Within a row there are 5 pairs of boxplots for the MLE and 5 pairs for the JE. The left member of each pair corresponds to pooling all faults together into one group. The right member corresponds to grouping faults according to type. Each pair represents a different issue mix $(N_{\mathrm{A}}, N_{\mathrm{B}})$ ranging from $(100, 0)$ on the left to $(0, 100)$ on the right. Arrows on the top or bottom of some of the boxplots indicate that a portion of the estimation errors were greater than 50 in magnitude.

Table 2 reports Ef_0 for the 20 study points. Table 3 reports median estimation errors, med$(f_0 - \hat{f}_0)$, for the MLE. The upper left entry in each cell of the table corresponds to pooling faults whereas the lower right entry corresponds to grouping them by type. Table 4 is for the JE and has the same format as Table 3. Standard errors for the entries in Table 3 (based on an estimate of the probability density at the median) are less than 1.1 faults and less than 4% of Ef_0 except in cell $(2, 5)$ in which case f_0 is nearly always zero and $\hat{f}_0$ nearly always rounds to zero. The entries in Table 4 have standard errors of less than 0.7 faults and less than 15% of Ef_0. Using medians rather than means to measure location allows an estimator to be considered well centered even if its distribution has a long right tail that greatly affects the mean. In particular, the MLE is infinite with positive probability and hence the theoretical mean estimation error is also infinite. The median error, however, is sometimes near zero. In the discussion of Figure 5 to follow we will refer to values from Tables 2 and 3 without reference.

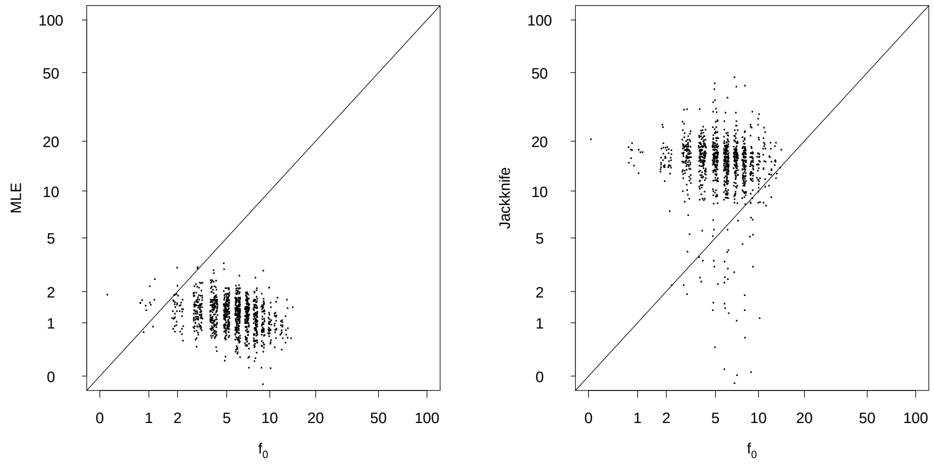

Figure 3: **Simulation Results – Two Fault Types Estimated Together.** Plots of $\hat{f}_0$ (MLE on left, jackknife on right) versus f_0 for two fault types $(N_\mathrm{A}, N_\mathrm{B}) = (75, 25)$, Case 2 $\mathbf{p}_\mathrm{A} = (.4, .4, .4, .4, .4)$ and $\mathbf{p}_\mathrm{B} = (.9, .9, .9, .9, .9)$, and fault types **not** distinguished by the estimators. Both the MLE and the JE are substantially biased.

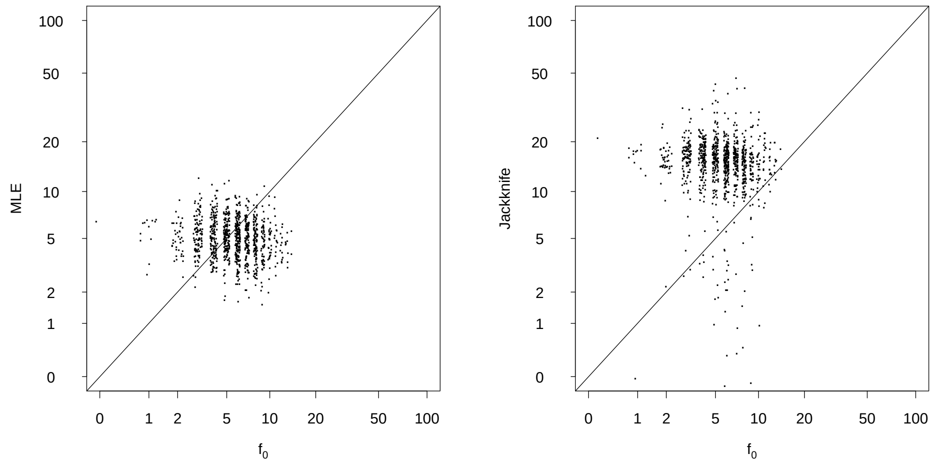

Figure 4: **Simulation Results – Two Fault Types Estimated Separately.** Plots of $\hat{f}_0$ (MLE on left, jackknife on right) versus f_0 for two fault types $(N_\mathrm{A}, N_\mathrm{B}) = (75, 25)$, Case 2 $\mathbf{p}_\mathrm{A} = (.4, .4, .4, .4, .4)$ and $\mathbf{p}_\mathrm{B} = (.9, .9, .9, .9, .9)$, and fault types estimated separately. Note the improvement in the median amount of underestimation for the MLE between Figure 3 and here: the median has shifted from -4.5 to -.6. The JE remains a poor estimator.

Table 2: Expected number of undiscovered faults for the 20 study points simulated.

$(N_\mathrm{A}, N_\mathrm{B}) =$	$(100, 0)$	$(75, 25)$	$(50, 50)$	$(25, 75)$	$(0, 100)$
$\mathbf{p}_\mathrm{A} = (.1, .1, .1, .1, .1)$ $\mathbf{p}_\mathrm{B} = (.6, .6, .6, .6, .6)$	59.1	44.5	30.0	15.5	1.02
$\mathbf{p}_\mathrm{A} = (.4, .4, .4, .4, .4)$ $\mathbf{p}_\mathrm{B} = (.9, .9, .9, .9, .9)$	7.8	5.8	3.9	1.9	0.001
$\mathbf{p}_\mathrm{A} = (.1, .1, .3, .8, .8)$ $\mathbf{p}_\mathrm{B} = (.8, .8, .3, .1, .1)$	2.27	2.27	2.27	2.27	2.27
$\mathbf{p}_\mathrm{A} = (.1, .3, .3, .3, .3)$ $\mathbf{p}_\mathrm{B} = (.8, .2, .2, .2, .2)$	21.6	18.2	14.9	11.5	8.19

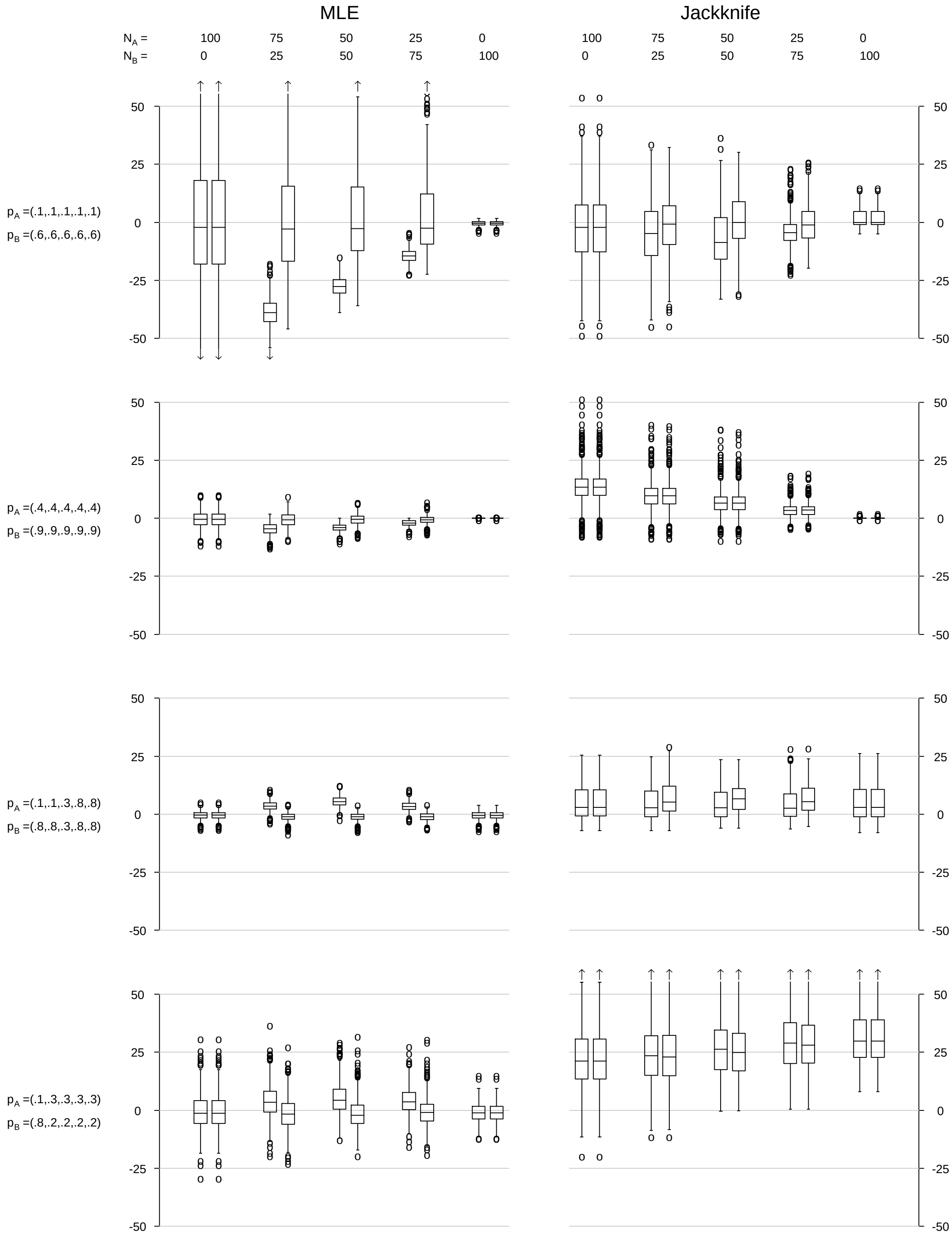

Figure 5: **Box Plots of $\hat{f}_0 - f_0$ for all Simulated Data Sets.** The left member of each pair of boxplots corresponds to pooling all faults together. The right member corresponds to grouping faults according to type. Grouping has little effect on the JE. For the MLE, grouping typically brings the median error closer to zero while increasing the variability somewhat. For example, in row 1 note how the "grouped" boxes for the MLE are better centered but taller than the "pooled" ones.

Table 3: Median number of faults overestimated by the MLE when faults are pooled (upper left corners) and grouped by type (lower rights).

$(N_{\mathrm{A}}, N_{\mathrm{B}}) =$	$(100, 0)$	$(75, 25)$	$(50, 50)$	$(25, 75)$	$(0, 100)$
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .1, .1, .1)$ $\mathbf{p}_{\mathrm{B}} = (.6, .6, .6, .6, .6)$	-2.2 -2.2	-38.9 -2.9	-27.7 -2.7	-14.6 -2.6	-0.5 -0.5
$\mathbf{p}_{\mathrm{A}} = (.4, .4, .4, .4, .4)$ $\mathbf{p}_{\mathrm{B}} = (.9, .9, .9, .9, .9)$	-0.5 -0.5	-4.5 -0.6	-4.0 -0.6	-2.0 -0.6	0.0 0.0
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .3, .8, .8)$ $\mathbf{p}_{\mathrm{B}} = (.8, .8, .3, .1, .1)$	-0.4 -0.4	3.5 -1.0	5.4 -1.1	3.4 -1.0	-0.5 -0.5
$\mathbf{p}_{\mathrm{A}} = (.1, .3, .3, .3, .3)$ $\mathbf{p}_{\mathrm{B}} = (.8, .2, .2, .2, .2)$	-1.3 -1.3	3.5 -1.6	4.3 -2.1	3.6 -1.0	-1.1 -1.1

Table 4: Median number of faults overestimated by the JE when faults are pooled (upper left corners) and grouped by type (lower rights).

$(N_{\mathrm{A}}, N_{\mathrm{B}}) =$	$(100, 0)$	$(75, 25)$	$(50, 50)$	$(25, 75)$	$(0, 100)$
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .1, .1, .1)$ $\mathbf{p}_{\mathrm{B}} = (.6, .6, .6, .6, .6)$	-2.3 -2.3	-4.8 -0.8	-8.8 -0.1	-4.5 -1.2	0.0 0.0
$\mathbf{p}_{\mathrm{A}} = (.4, .4, .4, .4, .4)$ $\mathbf{p}_{\mathrm{B}} = (.9, .9, .9, .9, .9)$	13.4 13.4	9.7 9.7	6.6 6.6	3.4 3.5	0.0 0.0
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .3, .8, .8)$ $\mathbf{p}_{\mathrm{B}} = (.8, .8, .3, .1, .1)$	3.0 3.0	2.8 5.3	2.8 6.7	2.6 5.4	3.0 3.0
$\mathbf{p}_{\mathrm{A}} = (.1, .3, .3, .3, .3)$ $\mathbf{p}_{\mathrm{B}} = (.8, .2, .2, .2, .2)$	21.2 21.2	22.9 22.9	24.8 24.8	28.1 28.1	29.8 29.8

4.1 Study points with one fault type

In Figure 5 the boxplot pairs for study points with a single type of fault are positioned along the left, center and right margins. In this subsection we focus on these 16 pairs of boxplots. Naturally the pooled (left) and grouped (right) boxplots in each pair are identical because all the faults are of the same type. It is interesting, however, to compare the MLE to the JE and to observe how estimation errors change with the discovery probabilities $\mathbf{p}_{\mathrm{A}}$ or $\mathbf{p}_{\mathrm{B}}$.

Because these study points conform to the assumptions of the MLE, it is not surprising that the MLE usually outperforms the JE. In fact, the JE does quite poorly for many of these cases. For example, in row 4 with $N_{\mathrm{A}} = 100$ the JE usually overestimates f_0 by more than 20 faults, while errors from the MLE have a median of -1.3. The variability of JE errors is also much greater.

The one study point (with a single fault type) where the JE does better than the MLE is in the top row with $N_{\mathrm{A}} = 100$. Here both the MLE and the JE give estimation errors with medians near zero but the errors from the JE are less variable.

We make the following generalizations from Figure 5 about cases with a single type of fault. The MLE has a distribution which is centered slightly below that of f_0. In our study the median estimation error ranged from -2.2 to 0. The JE on the other hand can severely overestimate f_0—often by more than a factor of 2. Estimation errors from the MLE also tend to be less variable than those of the JE. The reverse can also be true, however, especially when discovery probabilities are small.

4.2 Study Points with two fault types

In Figure 5 the 24 pairs of boxplots representing study points with two types of faults are those that do not border a margin—left, center or right. We now discuss these.

Looking over the 12 pairs of boxplots for the MLE immediately shows that the MLE can be severely biased when faults of different types are pooled and most of the bias can be removed if faults can be grouped according to type. The JE is much different. It is only slightly affected by mixed fault types. It can be severely biased regardless of whether faults are pooled or grouped.

Scanning across row 1 shows the effect of mixing an ever larger fraction of relatively easy to find Type B faults (60% discovery probability per reviewer) with hard to find Type A faults (10% probability). Mixing just 25 Type B faults with 75 of Type A causes the MLE (without grouping) to severely underestimate f_0. In this case grouping affects the MLE in two major ways. It brings the median estimation error up to near zero but also dramatically increases its variability. This nearly threefold increase in the interquartile range is due mostly to the difficulty in estimating the number of undiscovered Type A faults. Inspecting the other study points with only one fault type (see appropriate portions of Table 3 and Figure 5 in Section 4) shows that the distribution of the MLE is characteristically well centered over that of f_0. The estimation errors have medians ranging from -2.2 to 0.

Variability of the estimation errors depends, to a large extent, on the specific fault discovery probabilities. In one study point both f_0 and $\hat{f}_0$ were virtually always equal to 0 so there was almost never an estimation error. In another study point the interquartile range of the errors was 36 faults—more than half the median of 59.1. [The interquartile range is the difference between the 1st and 3rd quartiles (25th and 75th percentiles) and is a measure of variability.]

By contrast, the JE performs relatively well throughout row 1. It is only slightly affected by mixing Type A and B faults. In row 1 the JE usually performs better than the MLE (either with or without grouping).

Some of the MLEs in row 1 of Figure 5 were infinite. When no fault is found by more than one reviewer, the data indicates that there are many difficult to find faults. The more faults there are and the more difficult they are to find, the more likely it is that no fault will be found by two or more reviewers. The MLE is forced to be infinite. Thus, for a small group of difficult to find faults capture-recapture data often provides little information on the size of the group. If faults are to be grouped, it is important that in each group it is likely for several faults to be found by at least two reviewers.

Infinite estimates occur only for the MLE and in our simulation they appeared only in row 1 of Figure 5. They could have appeared in any row but the probability of this is extremely small in rows 2, 3 and 4.

Row 2 of Figure 5 shows the effect of mixing easy to find Type B faults (90% discovery probability per reviewer) with moderate Type A faults (40% probability). The MLE is affected in the same direction as in row 1—adding easy to find faults to the mix causes the MLE to underestimate f_0. Now the effect is smaller, though, because in row 2 most of the faults are found.

The JE in row 2 does poorly. It overestimates f_0 too often and is more variable than the MLE.

To understand why mixing fault types can bias the MLE so severely when reviewers are identical, consider an extreme case with several identical reviewers where half of the faults are easy to find (p near 1) and half are difficult (p near 0). Then each reviewer will discover nearly all of the faults that are found. The model for the MLE entertains only 1 fault type. Under this model, if all reviewers find essentially the same faults, then each reviewer is judged to have found nearly all the faults that exist. The, estimated number unfound is near 0 when in fact about half of the faults were not found. The effect is similar in less extreme cases. Mixing very different fault types is worse than mixing similar ones.

Rows 1 and 2 of Figure 5 showed that mixed fault types can cause the MLE to underestimate $f_0$when reviewers are identical. In rows 3 and 4 mixed fault and mixed reviewers types cause the MLE to *over*estimate f_0. In row 3 two reviewers specialize in finding Type A faults and two others specialize in Type B faults. When, for example, there are 50 faults of each type, the MLE nearly always overestimates f_0. The JE has a better centered distribution but it is also much more variable. Grouping faults by type improves the MLE and has a small adverse effect on the JE.

Finally, in row 4 all faults are difficult for all reviewers to find except that one reviewer finds one type of fault easily. Here the MLE is affected, but not severely, by the mix of fault types. Nevertheless, grouping faults by type brings the median estimation error closer to zero and only slightly increases its variability. The JE in this case is severely biased regardless of the fault mixture and whether faults are pooled or grouped.

We make the following generalizations from Figure 5 about estimating f_0 when there are two types of faults. A mixture of fault types can cause the MLE to either over- or under-estimate f_0 depending on the number and types of faults and reviewers. The effect can be severe, easily producing estimates as small as 20% of f_0 or as large as 4 times f_0. Grouping faults into types generally brings the median estimation error closer to zero but in some cases can dramatically increase the variability. If grouping is used, it is important to choose groups so that several faults in each group will be found by at least 2 reviewers. This will help to control the increased variability associated with grouping. The JE is not greatly affected by a mixture of fault types. Recall that the JE was *derived* to handle fault heterogeneity. However, the JE's nominal (single fault type) performance can be so poor that its performance with mixed fault types matters little.

4.3 Confidence bound comparisons

4.3.1 Study Points with 1 fault type

Tables 5 and 6 give coverage probabilities for Wald, Likelihood ratio and Jackknife upper confidence bounds for f_0. The bounds are constructed to have 95% nominal coverage. Table 5 presents coverages for all 20 study points but including only the Type A faults. Table 6 gives coverages for Type B faults. We include only a single type of fault in these tables, making no attempt to study cases where two types of faults are pooled because the MLE performs so poorly with mixed fault types that coverage probabilities matter little. The next subsection considers cases with 2 fault types when type *is* taken into account. Even in some of the single fault study points, however, the JE can severely over-estimate f_0. In these cases JE bounds are much too high so the coverage is near 100%.

Using the binomial variance typical standard errors for the entries in Tables 5 and 6 are 1.3%, 0.7% and 0.3% for entries near 80%, 95% and 99% respectively.

The tables show that the coverage of Wald bounds is often too low (typically 75% to 85%.) Likelihood ratio bounds have true coverage probabilities closer to the nominal value of 95% (typically 93% to 99.5%.) Likelihood ratio bounds, however, are often too conservative; many of the coverage probabilities exceed 99%. In rows 2 and 4 of Table 5 and row 4 of Table 6 the coverages for Jackknife bounds are too high because of the jackknife's large positive bias. In the other rows, where the bias is not so severe, coverage is too low (typically between 70% and 90%) like the coverage for Wald bounds.

4.3.2 Study Points with 2 fault types

Table 7 gives coverage probabilities for Wald, Likelihood ratio and Jackknife upper 95% confidence bounds for f_0. Coverages for the three confidence bounds when faults are grouped by type are similar to coverages for single fault types. Coverage for Wald bounds is usually too low—typically from 80% to 90%. Coverage for Likelihood ratio bounds is usually too high making them conservative. Coverage for Jackknife bounds is greatly affected by the bias of the JE. It is only 58% in cell (1,5) and at or near 100% in every cell of row 4.

5 Summary

Capture-recapture methods have recently been used to estimate the number of faults remaining in software design documents. Two possible estimators are the maximum likelihood estimator (MLE) and the jackknife estimator (JE). This work presents simulation results that show the following.

- The JE can greatly overestimate f_0 when reviewers, detection probabilities differ. For example, suppose a document has 100 faults. If one reviewer has a .8 probability of finding any given fault and each of 4 other reviewers has a .2 probability, then the JE is usually more than 3 times f_0.
- Both the JE and the MLE can estimate poorly when some faults are easy to find and others are difficult. For example, if a document has 75 faults with .4 (per reviewer) discovery probabilities and 25 faults with .9 probabilities, then the median of the MLE is 1.3 when 5.8 remain undiscovered on average. The median of the JE is 15.7 or about 2.6 times the expected remaining.
- If faults can be categorized into types then estimates can be formed for each of the types separately. This greatly reduces (but does not eliminate) the bias of the MLE.

A final finding of the simulation concerns upper confidence bounds when all faults are of one type. Bounds based on likelihood ratio tests tend to have coverage probabilities closer to their nominal levels than bounds based on either Wald statistics or the JE.

Three general recommendations can be drawn from these findings. First, if using the MLE, try to divide faults into a small number of groups so that within each group it is reasonable that a given reviewer will have a fixed probability of finding any fault. Groups formed in this way should be large enough that some faults in each group are discovered by more

Table 5: Attained coverage percentages for Wald (upper left corners), Likelihood Ratio (centers) and Jackknife (lower rights) upper 95% confidence bounds for Type A faults

$N_{\mathrm{A}} =$	100	75	50	25
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .1, .1, .1)$ $\mathbf{p}_{\mathrm{B}} = (.6, .6, .6, .6, .6)$	86.2 99.5 88.5	86.3 99.0 87.0	84.7 98.6 85.1	78.7 96.8 69.8
$\mathbf{p}_{\mathrm{A}} = (.4, .4, .4, .4, .4)$ $\mathbf{p}_{\mathrm{B}} = (.9, .9, .9, .9, .9)$	89.6 99.8 99.9	86.6 99.6 99.5	84.6 98.4 98.0	81.2 95.6 97.2
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .3, .8, .8)$ $\mathbf{p}_{\mathrm{B}} = (.8, .8, .3, .1, .1)$	85.9 98.1 80.6	83.9 96.6 75.1	81.3 93.4 69.1	77.1 91.3 88.2
$\mathbf{p}_{\mathrm{A}} = (.1, .3, .3, .3, .3)$ $\mathbf{p}_{\mathrm{B}} = (.8, .2, .2, .2, .2)$	88.9 99.7 99.7	88.7 99.6 99.7	85.5 98.9 99.7	82.6 96.7 97.8

Table 6: Attained coverage percentages for Wald (upper left corners), Likelihood Ratio (centers) and Jackknife (lower rights) upper 95% confidence bounds for Type B faults

$N_{\mathrm{B}} =$	25	50	75	100
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .1, .1, .1)$ $\mathbf{p}_{\mathrm{B}} = (.6, .6, .6, .6, .6)$	79.2 93.8 85.7	80.9 95.2 89.6	83.7 95.7 66.8	83.8 96.5 57.7
$\mathbf{p}_{\mathrm{A}} = (.4, .4, .4, .4, .4)$ $\mathbf{p}_{\mathrm{B}} = (.9, .9, .9, .9, .9)$	100.0 100.0 1.4	99.9 99.9 2.1	100.0 100.0 3.7	99.8 99.8 3.2
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .3, .8, .8)$ $\mathbf{p}_{\mathrm{B}} = (.8, .8, .3, .1, .1)$	76.9 92.6 88.6	81.5 93.5 73.4	81.9 96.9 76.8	85.7 97.6 79.2
$\mathbf{p}_{\mathrm{A}} = (.1, .3, .3, .3, .3)$ $\mathbf{p}_{\mathrm{B}} = (.8, .2, .2, .2, .2)$	72.9 90.1 99.6	80.7 96.9 100.0	86.0 98.0 100.0	86.3 99.2 100.0

Table 7: Attained coverage percentages for Wald (upper left corners), Likelihood Ratio (centers) and Jackknife (lower rights) upper 95% confidence bounds for when faults are grouped by type.

$(N_{\mathrm{A}}, N_{\mathrm{B}}) =$	$(100, 0)$	$(75, 25)$	$(50, 50)$	$(25, 75)$	$(0, 100)$
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .1, .1, .1)$ $\mathbf{p}_{\mathrm{B}} = (.6, .6, .6, .6, .6)$	86.2 99.5 88.5	86.0 98.8 87.9	83.9 97.7 88.6	77.7 84.8 78.6	83.8 96.5 57.7
$\mathbf{p}_{\mathrm{A}} = (.4, .4, .4, .4, .4)$ $\mathbf{p}_{\mathrm{B}} = (.9, .9, .9, .9, .9)$	89.6 99.8 99.9	86.6 99.6 99.5	84.6 98.4 98.0	81.2 95.6 97.5	99.8 99.8 3.2
$\mathbf{p}_{\mathrm{A}} = (.1, .1, .3, .8, .8)$ $\mathbf{p}_{\mathrm{B}} = (.8, .8, .3, .1, .1)$	85.9 98.1 80.6	78.4 95.1 95.6	78.1 93.6 92.1	77.6 95.1 95.5	85.7 97.6 79.2
$\mathbf{p}_{\mathrm{A}} = (.1, .3, .3, .3, .3)$ $\mathbf{p}_{\mathrm{B}} = (.8, .2, .2, .2, .2)$	88.9 99.7 99.7	86.6 99.6 99.9	85.1 99.4 100.0	85.1 99.3 100.0	86.3 99.2 100.0

than 1 reviewer. Otherwise, the MLE will be infinite. Second, recognize that even after grouping, the MLE tends to estimate too low—from 7% to 62% in the simulated study points. Third, use confidence bounds based on the likelihood ratio test statistic. They tend to have more accurate coverage probabilities.

Grouping faults into types has the effect of better satisfying the assumptions of the MLE. As we anticipated, grouping faults by type does not improve the JE. The equivalent technique for the JE would be to group reviewers by speciality. This is a possible area for further study. With typically only 5 to 7 reviewers, however, dividing them into "homogeneous" groups could be difficult.

Acknowledgement

We wish to thank Mark Ardis, Steve Eick, Willa Ehrlich, Diane Lambert and Dave Weiss for helpful comments on earlier versions of this manuscript.

References

[1] Barry W Boehm. *Software Engineering Economics.* Prentice Hall, Englewood Cliffs, NJ, 1981.

[2] K P Burnham and W S Overton. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65:625–633, 1978.

[3] Stephen G Eick, Clive R Loader, M. David Long, Scott A Vander Wiel, and Lawrence G Votta. Estimating software fault content before coding. In *Proceedings of the 14th International Conference on Software Engineering*, pages 59–65. Association for Computing Machinery, May 1992.

[4] Stephen G Eick, Clive R Loader, M. David Long, Scott A Vander Wiel, and Lawrence G Votta. Capture-recapture and other statistical methods for software inspection data. In *Computing Science and Statistics: Proceedings of the 25th Symposium on the Interface.* Interface Foundation of North America, 1993. To Appear.

[5] Watts S. Humphery. *Managing the Software Process.* Addison-Wesley Publishing Co., 1989. Reading, Massachusetts.

[6] G A Miller. The magic number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.

[7] David L Otis, Kenneth P Burnham, Gary C White, and David R Anderson. *Statistical Inference from Capture Data on Closed Animal Populations.* Wildlife Monographs. The Wildlife Society, Bethesda, MD, 1978.

[8] Gary C White, David R Anderson, Kenneth P Burnham, and David L Otis. Capture-recapture and removal methods for sampling closed populations. Technical report, Los Alamos National Laboratory, Los Alamos, NM, 1982.

List of Tables

List of Figures