

文章编号: 2096-1472(2016)-03-41-04

一种基于 p -值修正法的软件错误数估算的可靠算法

黄焯恒, 孙鉴锋, 王建新

(北京林业大学信息学院, 北京 100083)

摘要: 软件内部错误数估算是降低软件工程风险、保证软件工程顺利实施的重要途径。在当前软件测试中, 项目管理人员采用传统的估算策略, 但是估算效果往往受制于测试人员独立性等诸多因素的干扰。其中, 具有相似背景的测试人员对于测试过程会有相似认知, 这种不独立性使软件错误数的模型估算值往往比真实值偏低。类似潜在因素在软件错误数估算中大量存在, 降低了估算准确性。因此, 本文基于传统的Hyman估算法提出了一种改进的度量方法(p -值修正法), 该方法能有效排除组间人员相关性对Hyman模型的干扰, 能在很大程度上提高估算的准确性。同时, 该方法易于用户理解、简便易行、可靠性高, 可降低软件工程风险, 对决策人员有很大帮助, 适合普遍推广。

关键词: 软件测试; 软件错误估算; Hyman估算法; p -值修正法

中图分类号: TP311.5

文献标识码: A

A Reliable Algorithm for Software Error Estimation Based on Revision Method of p -value

HUANG Zhuoheng, SUN Jianfeng, WANG Jianxin

(School of Information, Beijing Forestry University, Beijing 100083, China)

Abstract: Estimating the inner error number of software is an important means of decreasing software engineering risks and making software engineering implemented successfully. In current software testing, project managers employ traditional estimation strategy, and the estimation effect is affected by the factors such as the independence of the test personnel. Testers of similar background may have similar cognition of the testing process, and such independence makes the estimated number of errors smaller than the real number. Moreover, such potential factors exist a lot in error estimating, hence destroying the estimation accuracy. Therefore, an improved estimation method, p -value revision method, is proposed in this paper, which is able to remove the intervention by the inter-group correlations of the testers, promoting the estimation accuracy. Meanwhile, this method is easy to be understood by users, is ready to be implemented, and is highly reliable. It is capable of decreasing software engineering risks, is helpful for decision makers, and is suitable for wide use.

Keywords: software testing; software error estimation; hyman estimation method; revision method of p -value

1 引言(Introduction)

当前, 随着互联网时代的蓬勃发展, 各类软件层出不穷, 数不胜数的软件在互联网这趟高速列车上蜂拥而至。并且, 随着用户需求的不断增加, 软件设计和开发过程中的复杂性不断增高。为了确保软件质量^[1,2], 应该开展更加行之有效的软件测试^[3,4]。计算机之父图灵认为软件测试是软件正确性确认的一种方法。因此, 软件测试应尽力找出软件设计的失败与不足之处, 并加以纠正, 确保软件最终设计目标的实现, 以此不断地提高软件质量。1983年, IEEE对软件测试给出了如下的定义: 使用人工或自动手段来运行和测定某个系统的过程, 其目的在于检验它是否满足规定的需求或是确定预期的结果与实际的结果之间的差别^[5]。但在测试之前需使用良好的软件错误估算方法^[6], 帮助测试人员发现错误, 使预期结果和实际结果之间的距离不断缩小, 是软件工程过程有更

好的可控性。

软件中错误数量的估算不准, 对软件质量的改善就越不确定, 人们发现一个错误越晚, 修正的难度越大, 造成的损失也就越大。所以软件的测试工作越早开始越好, 软件错误数的估算越准确越好。因而, 良好的错误估算模型对于提升软件质量犹如“银弹”般至关重要。

回首过往, 自软件行业诞生以来, 测试人员寻找良好的估算错误方法脚步一直没有停下来。从20世纪70年代开始到80年代中期是软件可靠性建模技术的快速发展时期, 这段时间产生了上百种的软件可靠性模型。1972年Mills将种子模型(Seeding-Model)运用到了软件错误数估算中^[7]。在过去, 种子模型是估计软件错误数量的重要手段之一, 该方法模仿估算池塘中鱼尾数的方式, 在开始排错以前被测软件中含有的错误数 x 。在不让排错程序人员知道的情况下, 在程序中置

入 y 个错误(相当于为 y 条鱼尾作标记后放回池中)。经过一段时间的软件测试,所测得的软件错误可以分成两类,一类是属于置入的错误 X ,另一类是非置入的错误 Y 。通过构造其比例关系式,来成功的估算出全部的错误数量。该方法具有简单直观的优点。但是该方法全过程不能自动实现,并且,无法保证人为置入的软件错误的规律与软件内实际存在的软件错误的规律一样或相似。后来,Hyman改进了Mills的估算错误的方法^[8],该方法可以运用蒙特卡罗(Monte-Carlo)模拟方法^[9]快速计算来解决这一问题,因而,这一模型被软件行业广泛应用。近来,苗洋^[7]通过对Hyman估算公式的变形,构造出估错数量与时间之间的函数,但是,该方法由于其较难操作,很难应用于实际的软件估错当中。

事实上,软件测试中估算错误数量的过程会受到诸多外界因素的干扰。比如,当两组测试人员不独立且高度相关时(这也是经常出现的情形),其测得的相同错误数会偏低,从而使估算值比真实值偏高。并且,这种相关性越强,估算值的偏离程度越大。

为了尽可能缩减这一偏差,改善Hyman估算方法的质量,我们提出一种相似度系数(p -值修正法)来改善估算值的准确性。该方法通过度量测试人员之间的相关程度,运用“惩罚”的思想,修正估算结果,对软件中的错误数做出更加精准的判断。为此,我们在文中引入对Hyman法的概率推导过程^[6],帮助读者更好的理解 p 值修正法的整体思想。

2 估算软件中错误数量的方法(Method of estimating software error number)

2.1 Hyman估算法

假设 A 组和 B 组两组测试人员互相独立地对某个软件进行测试,我们记 A 组人员和 B 组人员测得的错误数分别为 i 和 j 个,两组测试人员共同测试出的错误数为 k ,软件错误数的估算值 $\hat{N}$ 与这三个量的关系如下:

$$\hat{N} = \frac{i \times j}{k} \quad (1)$$

式(1)即为著名的Hyman估错方法表达式。该模型简单易懂,但在排错人员进行软件测试之前,通常要做如下几个假设:

(1)两组测试人员之间是相互独立的,并且在测试过程中没有相互交流。

(2)排错时,不再引入新的错误。

(3)排错的过程随机进行。

(4)在排错过程中错误特性不变。

(5)所有软件缺陷的触发概率都是相等的。

如果上述某条假设不成立,均会在一定程度上降低最终结果的准确性。下面我们用概率模型的手段分析该问题。

一方面,由于实际生产生活中,估算软件错误数目带有未知性和不确定性,因此,我们可以将这一过程抽象为一种概率问题;另一方面,这为了更好地帮助理解本文提出的优化方法,引入针对Hyman法问题求解的概率模型,记事件 Z 为“ A 组人员和 B 组人员检测出的相同错误的数量恰好为 n ”,所以可有:

$$P(Z) = \frac{C_i^k C_{n-i}^{j-k}}{C_n^k} \quad (2)$$

其中 n 为估算个数,也就是说我们希望找到最佳的 $n = \hat{N}$,

令 $f(n) = P(Z)$,研究函数 $f(n)$ 的单调性。若 $\frac{f(n)}{f(n-1)} > 1$,则函数单调递增;若 $\frac{f(n)}{f(n-1)} < 1$,则函数递减。

将(2)式带入 $\frac{f(n)}{f(n-1)} > 1$ 当中,推得

$$n < \frac{i \times j}{k} \quad (3)$$

即当 $n < \frac{i \times j}{k}$,函数 $f(n)$ 递增;当 $n > \frac{i \times j}{k}$,函数递减,即 $\hat{N} = \frac{i \times j}{k}$,这也证明了采用Hyman估错的可靠性。

但是,当两组测试人员具有较高的相关性时,Hyman方法会有较大的估算误差,如下面场景。

场景1:假设M公司的某款软件中存在100个错误, A 组和 B 组两组测试人员在对该软件中测试时,分别得到了 $i=20$ 和 $j=20$ 个错误。理想情况下,如果两组测试人员之间完全独立,则测出的共同错误数 k 的理论值为4,此时用式(1)估算的结果为100恰好与实际值相等;但如果两组测试人员之间不独立且有正相关性,则检测出的共同错误数 k 大于4,比如 $k=4$ 。此时根据公式(1)的估算值为40,与实际只有较大偏差。可以看出,独立性越差,偏差就越大。

经过对软件测试行业现状的分析和以往软件测试的经验,我们发现场景1中出现的测试人员的相关性问题会在很大程度上左右最终的估算结果,从而导致了较大的估算误差。人员相关因素与人员的背景密切相关。例如,长时间共事的两组测试人员,他们在纠错认知方面势必具有相似之处,这将在一定程度上破坏Hyman模型中的假设,使估算值向低偏

离真实值。

从场景1的结果也进一步看到，当A组和B组人员之间的相关性越小时，越能提高估算结果的精确性。然而，由于两组测试人员得到的共同测试错误的数量 已经是一个定数且不能改变，因此，我们需要找出一种策略去弥补这种不足。基于上述考虑，本文提出了 p 值度量法(用 p 值来惩罚相关程度)来改进Hyman测试法。

2.2 改进Hyman法(p -值修正法)

p 值修正法以人员之间的相似程度等一些外在因素为度量依据，能够在一定程度上提升软件错误数量估计值的准确度。决策人员可以根据测试人员之间的不同特性，来及时的更换 p 值(结合“专家意见法”^[10]对人员相关因素的相关程度进行打分)的大小，这将对最终估算结果的正确性起到巨大的帮助。

为了让当前软件错误数量的估算值 $\hat{N}$ 更加接近错误数量的真实值 n ，我们有必要重新确定一个估计值 $\hat{N}'$ ，它可以被如下定义

$$\hat{N}' = \frac{1}{(1-p)} \hat{N} \quad (4)$$

其中， $1 > p > 0$ ，观察(4)式，可以看到 p 值其实就是一个在0—1的惩罚系数，这个惩罚的思想便是 p 值修正法的核心思想，该思想旨在消除两组人员不独立造成的影响，在干扰和估值之间做一个平衡。使获得的新估算值较好的接近软件错误数的真实值。其算法流程如图1所示。

本文认为，在实际软件测试当中，测试人员的先决条件(人员相关程度)能够影响测试过程中测出的共同错误数量。这些先决条件包括测试人员的学习经历和工作背景等情况，具体而言，对于衡量这个先决条件，主要通过几个方面进行：

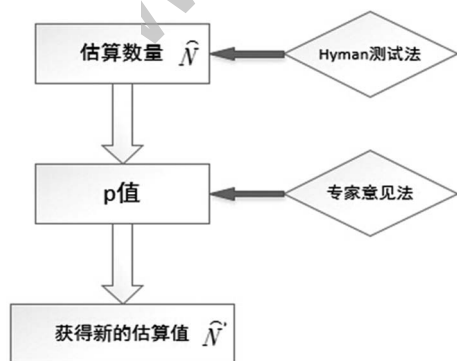


图1 p -值修正法流程图

Fig.1 Flow chart of revision method of p -value

第一，测试人员之间的相关性与测试人员之前所处的环境(地点)有关。例如，两组人员之中存在长期共事的情况，因为他们现今共同工作，对某一些类型的错误具有相同认识，使得测试结果中共同检测出的错误数量变多。再比如，两组人员中存在毕业同一院校的，他们之前的学习氛围相近，受到的同种技巧教育的影响偏深，也会提高对软件中某几类错误具有共同认识的可能性，从而也导致共同检测错误数量值变大。

第二，人员相关性施加于某项因素的时间相关。时间越长，相关性越大。

因此，我们立足于当前软件行业的发展现状和趋势，并模拟“专家意见这一过程”，总结出测试人员相关性的主要类别，分别是：

- (a)测试人员共事年限长短。
- (b)测试人员工作年限相似度。
- (c)测试人员毕业院校相似度。
- (d)测试人员学历相似度。

采用低、中、高作为相关程度的衡量标准，并且经由“专家意见法”，针对人员相关因素的相关程度进行打分，其取值如表1所示。

具体地，针对相关因素类型 Q_i 而言，分为情况 s_i (如表2所示)，并且每种情况 s_i 用相应的相关程度值 p_i 来权衡。表中的各类情况 s_i 随值 i 的增加，其相关程度也在增加。

表1 软件测试人员相关性的 p 值表

Tab.1 p -value table of correlation between testers

相关因素类型 Q_i	权重 W_i	相关程度值 p_i		
		低度	中度	高度
Q_1	0.4			
Q_2	0.3			
Q_3	0.2	0.1	0.4	0.8
Q_4	0.1			

注： Q_1 代表测试人员共事年限长短， Q_2 表示测试人员从业年限相似度， Q_3 表示测试人员毕业院校相似度， Q_4 表示测试人员学历相似度； W_i 对应相关因素类型 Q_i 的权重； p_i 代表测试人员相关程度高度相关， p_2 代表中度相关， p_3 代表低度相关。

本模型规定，在评价测试人员相关性时，如果组内存在相关因素类型 Q_i 的复合情况时，只保留相关程度值最大的一种，这称为复合情况保留策略。相应地， p_i 为 s_i 的对应值。

表2 相关类型对应的各类情况

Tab.2 Correlation types of different situation

相关因素类型 Q_i	情况 s_1	情况 s_2	情况 s_3
Q_1	0—1年	2—3年	4年或以上
Q_2	3—4年	1—2年	1年或以下
Q_3	无关院校	相似院校	相同院校
Q_4	学历差2级或以上	学历差1级	学历相同

在上述讨论的基础上，我们将 p 值修正法应用于场景1描述的问题中。当 M 公司 A 、 B 两个小组各自测得的错误数和共同测得的错误数分别为20、20和10时，根据Hyman估算法得到了总错误数的估算值40，与真实值100有较大的偏差。

但通过对测试人员内部的评价，可获得 A 、 B 两个小组人员的信息。其中相关类型 Q_1 出现了情况 s_1 、 s_2 、 s_3 (两组人员中共事年限最长为5年，根据复合情况的保留策略，取最相关情况 s_3 ，后同)；相关类型为 Q_2 出现了情况 s_1 、 s_2 (两组人员中从业年限差最小为1.5年，取最相关情况 s_2)；相关类型为 Q_3 出现了情况 s_1 、 s_2 (两组人员中无共同院校毕业，但有相似院校毕业情况，取最相关情况 s_2)，相关类型为 Q_4 出现了情况 s_2 、 s_3 (两组人员中有同为硕士毕业的情况，即学历差为0，取最相关情况 s_3)。那么， $\{Q_1, Q_2, Q_3, Q_4\}$ 分别保留了情况 $\{s_3, s_2, s_2, s_3\}$ ，对应了相关程度值 $\{p_3, p_2, p_2, p_3\} = \{0.8, 0.4, 0.4, 0.8\}$ ，最终的 p 值为

$$p = W1 \times p3 + W2 \times p2 + W3 \times p2 + W4 \times p3 = 0.6$$

上述结果带入(4)式，得到 p 值修正法的估算结果为 $2.5\hat{n}=100$ 。可以看出，这个结果比较接近真实的软件错误数估算结果。

通过表1中的 p 值惩罚系数，能够将Hyman估算法的偏离值复原，得到的数据也就更加可靠。另一方面，通过“专家意见法”，不同公司可以根据自己的实际情况，调整表2的赋值，使得 p 值更加符合公司和企业的状况，降低因错误数估算偏离造成的软件工程风险。

3 结论(Conclusion)

软件测试是软件工程中至关重要的环节，而软件错误数

的估算是实施有效的软件测试的前提。估值过高或估值过低都会影响软件工程的工期，造成资源投入的不足或浪费。我们提出的 p 值修正法是对传统的Hyman估算法的改进，克服其因人员的相关性而造成的估值偏差。该方法通过简单的 p 值修正即可使估算值与真实值之间的偏差大幅下降，并具有简单易懂、实施方便和高效快捷等优点，为降低软件工程风险、保证软件工程顺利实施提供了良好的估算机制。

参考文献(References)

[1] 聂芸.基于CMMI的软件质量保证过程管理[J].软件工程师, 2015,18(8):10-11.

[2] Baker E R,Fisher M J.Software quality program organization[M]. Prentice Hall PTR,1998:115-145.

[3] 赵丽莉,金学军.软件性能测试面面观[J].软件工程师,2006 (11):40-42.

[4] 陈明.软件工程教程[M].北京:科学出版社,2002.

[5] IEEE Standard Glossary,IEEE Std 729-1983.

[6] 李世慧,惠韶文.计算机软件中错误数量的一种估算方法[J].佛山科学技术学院学报:社会科学版,1995,13(2):52-53.

[7] 苗扬.软件可靠性测试与评估方法的改进[D].上海交通大学, 2010.

[8] 刘继华.基于风险的Web应用软件测试方案研究与应用[D].太原理工大学,2006.

[9] 马海云,党建武.基于蒙特卡罗的软件测试技术的研究与实现[J].郑州大学学报(工学版),2007,28(4):55-58.

[10] 戴怡,等.面向数控系统可靠性评估的先验信息融合方法—专家打分法[J].天津职业技术师范大学学报,2012,22(3):6-8.

作者简介:

黄焯恒(1994—)，男，硕士生.研究领域：软件工程。

孙鉴锋(1991—)，男，博士生.研究领域：软件工程，数据挖掘。

王建新(1972—)，男，博士，教授.研究领域：软件测试，软件工程，数据挖掘。