

多重検定の補正が及ぼす影響について

天寄 聡介^{1,a)}

概要: ソフトウェア工学分野では、新たな提案や発見が偶然ではないことを示すために統計的検定を行うことが一般的である。本稿では、多重検定の補正が効果量や検出力に及ぼす影響について述べる。

1. はじめに

ソフトウェア工学分野では統計的検定が日常的に用いられる。例えば、ある尺度で計測された実験結果の値に統計的検定を適用することで提案手法が先行研究よりも有用であることを主張しようとする。統計的検定によって帰無仮説が棄却された場合、実際的な有用性を主張するために効果量を示すこともある [1], [2]。

工数予測モデルやソフトウェア不具合予測モデルに関する研究では、様々なパラメータを変化させつつ比較実験を行うため、多重検定の問題が生じる。統計的検定を行うのが1回であれば、あらかじめ設定した有意水準と p 値を比較して、帰無仮説が棄却されるか判断すれば良い。複数回統計的検定を行う場合には試行回数に合わせて有意水準（もしくは p 値）を補正して第一種の過誤を制御する必要がある。

多重検定に対する最も基本的な補正法が Bonferroni の補正法である。有意水準が 5% で統計的検定を 10 回行う場合、同手法による補正後の有意水準は $5/10 = 0.5\%$ となる。昨今、Bonferroni の補正法はソフトウェア工学分野で広く用いられている [2], [3]。一方で、補正法の適用が及ぼす影響について論文中で言及されることは少ない。本稿では、多重検定の補正が効果量や検出力へ及ぼす影響について述べる。

2. 多重検定の補正による影響

多重検定に対する補正法は、第一種の過誤を念頭に置いた手法である。統計的検定における第一種の過誤とは、帰無仮説が真である場合に帰無仮説が棄却されることである。有意水準 5% で有意とは、この確率が 5% 以内である

ことを指す。一方、帰無仮説が偽である場合に帰無仮説が棄却されないことを第二種の過誤と呼ぶ。第二種の過誤の確率を β とした時に、 $(1 - \beta)$ を検出力と呼ぶ。

有意水準と検出力は、サンプルサイズ及び効果量と密接な関係にある。対照実験のように実験計画を事前に行う場合、期待される効果量や要求される有意水準、検出力を元にサンプルサイズを決定する。例えば、有意水準と検出力が一定で効果量が小さい場合、より大きいサンプルサイズが必要となる。

Bonferroni の補正法を適用した場合、有意水準の値は小さくなる。その結果、検出力と効果量を一定に保つのであれば、サンプルサイズをより多くする必要がある。一方、サンプルサイズと効果量を一定に保つのであれば、検出力は低下する。

3. 実例

ここでは、我々の研究 [4] を題材として、多重検定の補正による影響を例示する。文献 [4] では、Moving Window 法と呼んでいる工数見積もり手法を従来法と比較している。比較実験では、ウィンドウサイズというパラメータの値を 101 通りに変化させて評価実験を行なっている。また、パラメータの値に応じてテストデータのサンプルサイズが 71 から 201 まで変化している。有意水準は 5% としてある。

Bonferroni の補正法を適用すると、有意水準は約 0.05% となる。文献 [1] を参考に、効果量には実用的に意味のある下限である 0.2、中程度を示す 0.5、大きな効果を示す 0.8 をそれぞれ想定する。二群の差の検定についての検出力分析のために、Python のパッケージ `statsmodels` の関数 `tt_solve_power` を用いた。同関数は論文中のノンパラメトリック検定ではなく t 検定を想定している。一般にノンパラメトリック検定の検出力は通常の統計的検定より低いため同関数の出力はやや楽観的であると考えられる。

図 1 に結果を示す。まず、効果量が十分大きい場合、どのウィンドウサイズでも検出力は 1.0 である。つまり、現

¹ 岡山県立大学
Okayama Prefectural University, Soja, Okayama 719-1197, Japan

^{a)} amasaki@cse.oka-pu.ac.jp

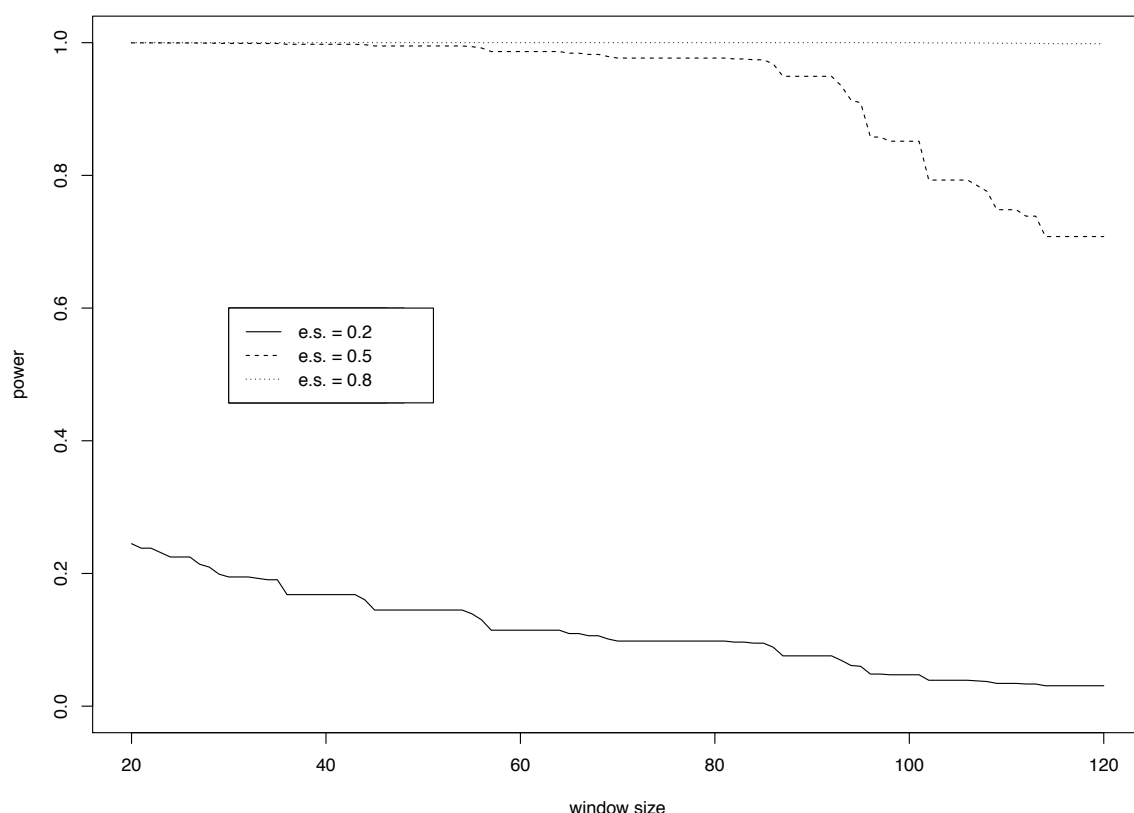


図 1 ウィンドウサイズ毎の検出力 (Bonferroni 補正後)

在のテストデータ数で十分であると言える。次に、効果が中程度の場合、ウィンドウサイズが 60 を越えたあたりから検出力が低下している。ウィンドウサイズが 120 のとき、検出力は 0.7 となった。検出力は 0.8 と設定する場合が多いため、今回の結果はやや低めであると言える。最後に、効果量が下限程度である場合、検出力は全てのウィンドウサイズで 0.8 を大きく下回っていた。ウィンドウサイズが 20 のときにテストケースが最も多いが、この場合でも検出力は 0.25 程度である。そしてウィンドウサイズが 120 のとき、検出力は 0.03 まで低下する。ちなみに、Bonferroni 補正を行う前であれば、効果量 0.2 の場合、ウィンドウサイズ 20 で 0.8 であり、ウィンドウサイズ 120 で 0.38 である。効果量 0.5 の場合、ウィンドウサイズ 120 で検出力 0.98 である。

以上の例から、比較している手法間の差が十分に大きくない場合、Bonferroni 補正法の導入によって検出力が低下し、有意な差が見逃されている可能性が高くなっていると考えられる。文献 [1] では 3 種類の類推法を比較しており、効果量は最大で 0.36 である。同文献では手法間で統計的有意差が確認されているが、上記の論文と同様の比較実験を行った場合には有意差が見逃されていた可能性も大きい。

4. まとめ

本研究では、多重検定の補正が効果量や検出力に及ぼす影響について考察した。検出力の低下を避ける方法として

は、実験方法の工夫や他の補正法 [5] を用いることなどが考えられる。後者について実際に適用して有用性を確認することが今後の課題である。

謝辞 本研究の一部は、日本学術振興会科学研究費補助金 (若手 (B):課題番号 15K15975) による助成を受けた。

参考文献

- [1] Shepperd, M. and MacDonell, S.: Evaluating prediction systems in software project estimation, *Inf Softw Technol*, Vol. 54, No. 8, pp. 820–827 (2012).
- [2] Minku, L. L. and Yao, X.: Ensembles and locality: Insight on improving software effort estimation, *Info Softw Technol*, Vol. 55, No. 8, pp. 1512–1528 (2013).
- [3] Lokan, C. and Mendes, E.: Investigating the use of moving windows to improve software effort prediction: a replicated study, *Empir Softw Eng*, pp. 1–52 (2016).
- [4] Amasaki, S. and Lokan, C.: On the effectiveness of weighted moving windows: Experiment on linear regression based software effort estimation, *Journal of Software: Evolution and Process*, Vol. 27, No. 7, pp. 488–507 (2015).
- [5] Gelman, A., Hill, J. and Yajima, M.: Why we (usually) don't have to worry about multiple comparisons, *Journal of Research on Educational ...*, Vol. 5, No. 2, pp. 189–211 (2012).