離散的に観測された欠陥データへのレイリーモデル適用について

野　中　　誠[†]

本稿では，離散的で測定間隔が不均一なソフトウェア欠陥データにレイリーモデルを適用する際の，欠陥データの処置方法を示す．この処置を施した実データにレイリーモデルを適用した結果，総欠陥数の予測精度が向上する例があった．一方で，そもそもレイリーモデルが適合しない例も観測された．

Applying the Rayleigh Model to Discrete Software Defect Data

Makoto Nonaka[†]

This paper explains a procedure for applying the Rayleigh software defect prediction model to discrete data which were obtained at different measurement intervals. The procedure showed better performance to predict the total nuber of defects in a practical dataset. On the other hand, a nonconforming dataset to the Rayleigh model was also observed.

1. はじめに

ソフトウェアライフサイクル全体にわたる欠陥数の予測に，レイリー（Rayleigh）モデルが適用できるという議論がある[1)2)3)]．レイリーモデルは，横軸に時刻という連続変数を，縦軸に時刻 t における欠陥発見数（または欠陥発見率）$f(t)$ をとった連続関数である．

しかし，実務的に測定できる欠陥は，週次や工程別などの離散的に観測されるデータであり，測定間隔も等しいとは限らない．たとえば，毎週 2 件の欠陥を観測した場合と，2 週間隔で 4 件を観測した場合では，総欠陥数は同じでも座標平面上では異なるプロットになるため注意を要する．さらに，工程別で欠陥発見数を測定した場合に，文献 1) が示すように各工程を等間隔で表すことが妥当であるのか議論の余地がある．

離散的に観測された欠陥データを，連続関数のモデルに当てはめる方法は単純かつ容易である．しかし，文献 1) や文献 2) ではその扱いについて何も言及していない．また，そうした処置を施した上で，実際の欠陥データにレイリーモデルを適用した報告は，筆者の知る限り示されていない．本稿では，離散的で，等間隔ではない欠陥データに対して，レイリーモデルを適用するときの具体的な処置方法を説明する．また，この処置を施した工程別の実データに対して，レイリーモデルを適用した際に得られる考察について述べる．

† 東洋大学
Toyo University

2. レイリーモデル

欠陥数を予測するレイリーモデルは次式である．

$$f(t) = K \times 2(t/c^2) \times e^{-(t/c)^2}. \qquad (1)$$

ここで，K は総欠陥数，c は形状パラメータであり，$f(t)$ が最大値となる時刻を t_m とすると $c = \sqrt{2}t_m$ が成り立つ．当該プロジェクトにおける時刻と欠陥発見数の組が少なくとも 3 つ与えられれば，非線形最小二乗法によりパラメータ K と c を推定できる．

なお，(1) 式で示したレイリーモデルは，ワイブル（Weibull）分布の確率密度関数

$$f(t) = m(t/c)^m \times e^{-(t/c)^m}/t \qquad (2)$$

において $m = 2$ としたレイリーモデルに，総欠陥数 K を掛けることで得られた式である．確率密度関数と横軸で囲まれた面積が 1 なので，これに K を掛けた (1) 式では，レイリーモデルの曲線と横軸で囲まれた面積が総欠陥数を示している．

3. 離散的に観測された欠陥データの処置方法

離散的に観測された欠陥データのヒストグラムを各欠陥除去工程の完了時刻に合わせて描く場合，時刻という連続変数を横軸にとると，それぞれの階級の幅は当然ながら不均一になる．また，前項で述べたとおり，レイリーモデルの曲線と横軸で囲まれた面積は総欠陥数に等しい．したがって，レイリーモデルを適用する場合には，各階級の柱の面積がそれぞれの欠陥除去工

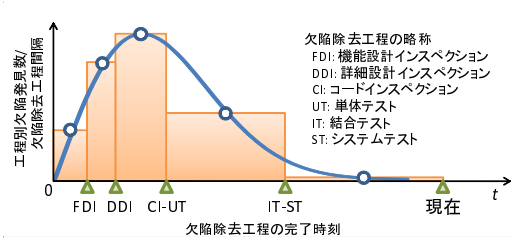

図 1　レイリーモデルと工程別欠陥除去数

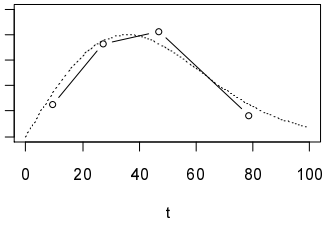

図 2　測定間隔を補正したデータへのレイリーモデル適用例 1

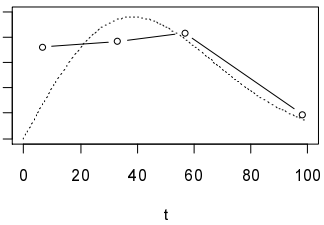

図 3　測定間隔を補正したデータへのレイリーモデル適用例 2

程での欠陥発見数に等しくなるようにヒストグラムを描く必要がある．すなわち，各階級の高さを，欠陥発見数を階級の幅で割った値（密度）とし，これを縦軸にとる必要がある．レイリーモデルのパラメータ K および c は，このようにして描かれたヒストグラムに対して適合するように求めるべきである．以上の説明を視覚的に表したものが**図 1** である．

この処置方法をフォーマルに表現すると次の通りになる．i 番目の欠陥除去工程において，その完了時刻を t_i，欠陥発見数を n_i とする．レイリーモデルのパラメータ推定に用いる i 番目のデータポイントを P_i とするとき，その座標は

$$P_i\left(t_{i-1}+\frac{t_i-t_{i-1}}{2},\frac{n_i}{t_i-t_{i-1}}\right) \qquad (3)$$

となる．ただし，$i>0, t_0=0$ である．

4. 実データへの適用

この処置方法を実際のソフトウェア開発データに適用し，レイリーモデルの適用性評価を試みた．ここでは，2005 年以降に開発されたソフトウェア製品の実データを用いた．この組織は品質管理や品質保証に関する長年の経験があり，成熟度が高い組織である．

図 2 と**図 3** に，実データにレイリーモデルを適用した 2 例を示す．図 2 では，レイリーモデルが高い精度で適合している様子が分かる．この例では，総欠陥数 K の予測誤差は 1.37% と極めて低く，精度の高い予測が行えている．一方，文献 1) の方法で横軸を等間隔とした場合の予測誤差は 7.57% であった．この例では，測定間隔を補正した効果が得られている．

しかし，図 3 では，レイリーモデルが適合しておらず，総欠陥数 K の予測誤差も 28.83% と大きい．このデータについて，ワイブル分布のパラメータ m を 1.1 から 2.0 まで 0.1 刻みで変化させたが，もっとも予測誤差が小さかったのが図 3 の推定モデルであった．初期での欠陥除去数があまりに高い場合には，レイリーモデルが適合しない可能性がある．

なお，本稿では 4 個のデータポイントを用いてパラメータ推定を行っており，推定値の標準誤差が非常に大きい．このことに留意しておく必要がある．

5. おわりに

近年の複雑なソフトウェア開発プロジェクトにはレイリーモデルが適合しないとの報告[4] もある．本稿が示した範囲では，レイリーモデルがうまく当てはまった例については，離散的な観測データの処置を行うことでモデルの適合性が高まることが示された．しかし，早期の欠陥除去が高い水準の場合には，そもそもレイリーモデルの適合性に問題があることが示された．

レイリーモデルなどの動的欠陥予測モデルは，当該プロジェクトのデータを用いる点において静的モデルにはない特性がある．動的モデルの特性を活かした品質予測技法について，本ワークショップで議論したい．

参考文献

1) Kan, S. H.: *Metrics and Models in Software Quality Engineering*, Addison-Wesley (2002).
2) Laird, L. M. and Brennan, M. C.: *Software Measurement and Estimation: A Practical Approach*, Wiley-IEEE Computer Society (2006).
3) Smith, C. and Uber, C.: Experience report on early software reliability prediction and estimation, *Proc. 10th Int'l Symp. Softw. Reliability Eng.*, pp. 282–284 (1999).
4) Yousef, A. H.: Analysis and enhancement of software dynamic defect models, *Int'l Conf. Networking and Media Convergence 2009(ICNM 2009)*, pp. 85–91 (2009).