

# 31SQiP 研究会 特別講義 レポート

作成日： 2015年 10月 16日

書記氏名： 齋藤 伸介

日時	2015年 10月 9日 (金) 10:00~12:00
会場	(一財)日本科学技術連盟・東高円寺ビル 2階講堂
テーマ	「メトリクスに統計手法を適用する意義、事例、ノウハウ」
講師名・所属	小池 利和 氏 (ヤマハ株式会社)
司会者	飯泉 紀子 氏 (株式会社日立ハイテクノロジーズ)
アジェンダ	<ul style="list-style-type: none"><li>・なぜソフトウェアメトリクス活用に統計手法が有効なのか</li><li>・事例を通じて統計手法適用のノウハウを解説</li></ul> 当たり前品質のマネジメント <ul style="list-style-type: none"><li>－開発終了後の事後分析事例</li><li>－開発中のメトリクス管理事例</li><li>－計画立案時の予測モデル活用事例</li></ul> 魅力品質のマネジメント <ul style="list-style-type: none"><li>－魅力品質マネジメントプロセスの概要</li><li>－ユーザビリティテストにおける活用事例</li><li>－音質の官能評価事例</li><li>－市場での魅力品質評価事例</li></ul>
アブストラクト	<p>ソフトウェアメトリクスは測定誤差が大きいと、データ指向による品質管理には向かないと考える人がいるかもしれません。特に理系出身で自然科学データを扱ってきた人はその傾向が強いようです。</p> <p>ですが、経済学や心理学など人間の活動を研究対象とする社会科学分野では測定誤差が大きいのは当たり前のこととして、その中で研究結果の信頼性を確保しなければなりません。そのような分野では測定誤差以上の差や効果が有るかを示すために統計手法が活用されています。ソフトウェアメトリクスは人間の活動を測定している側面が強く、まさに社会科学的データといえます。そのため、データを活用する上で統計手法は有用です。</p> <p>本講義では、ソフトウェアメトリクスに統計手法を適用する意義とその際のコツを事例を通して解説します。事例には当たり前品質としての欠陥を対象としたものだけに留まらず、ユーザビリティや音質といった魅力品質に関する内容も盛り込んでいます。</p>

今回の特別講演では「メトリクスに統計手法を適用する意義、事例、ノウハウ」と題してヤマハ株式会社 品質保証部の小池利和さんから、講演を頂きました。

小池さんは、社内において、長年、品質活動に取り組んでいると同時に、社外におかれても、SQiP 研究委員長、メトリクス演習コース主査として活動され、またソフトウェアメトリクスに関する著作も多数執筆されています。

今回の講演では、受講された皆さんが統計手法を現場活用できるように、と具体的な事例を多数交えて、非常にわかりやすい講演をしていただきました。ありがとうございました。

#### ■はじめに～本日のキーメッセージ

- ・推定誤差を常に意識することの重要性
- ・これからの品質管理は、製品の良さを評価できないと、競争に勝ち残れない。

#### ■ソフトウェアメトリクスの分析に統計手法を使う意義

- ・データ分析においては、2種類のエラーがある。
  - $\alpha$ エラー (何となく信じてしまう“あわて者”)
  - $\beta$ エラー (分析結果が信用できずアクションしない“ぼんやり者”)
  - どちらが正しいかは、“神のみぞ知る”、であれば、確率論をベースにした統計手法を用いる。
- ・ソフトウェアメトリクスは誤差が大きいから使えない？
  - 誤差はあって当たり前、だからこそ、統計手法を駆使して情報を読み取るべき。

#### ■ t 検定を用いたデータ分析とその事例

- ・社会科学分野で t 検定を用いた分析事例の紹介

事例1：夫婦の関係満足度および生活充実感における規定因の検討

事例2：恋人という間柄を意味する諸行為の記号学的分析

事例3：テスト形式の違いによる学習方略と有効性の認知の変容

→ 3つの研究例とも、2つの群に分けてデータを収集し、その平均に「偶然とは言えない差があるかどうか」を検証する「t 検定」という統計手法を用いて分析している。

- ・2群のデータの差を見る上では、「平均の差」「標準偏差」「データ数」を総合して判定する必要がある。その総合指標を「t 値」と呼ぶ。t 値の絶対値が大きいほど2群の差は大きいと言える。  
t 値の計算式  $t = d/s/\sqrt{n}$  : d:平均の差、s : 2群の標準偏差の加重平均、n : 2群のデータ数
- ・t 値が0に近いなら2群の平均はほぼ等しく、0から離れていれば2群は違いがある、と言える。その離れ具合を示す指標が「P 値」である。  
2群が十分離れている=等しくない=有意差あり、と判定するP 値の基準は慣例的に5%とされている。  
P 値が小さいほど、2群の平均は等しくない、と言える。

#### ■当たり前品質 (不具合) のマネジメント

事例1：開発終了後データを用いた分析事例

- ・分析の目的
- ・不具合密度 (システムテスト不具合数/新規開発行数) と不具合検出率 (システムテスト不具合数/システムテスト工数) のどちらが品質指標として妥当か、を検証する。
- ・分析結果
  - － 不具合密度と品質問題数の相関分析
    - それほど高くなかった。
  - － リリース後の「品質問題無し群」と「品質問題有り」の2群の平均の差を t 検定した。
    - P 値が 0.36、となり 5%より大きいので2群の平均値の差があると言えない。  
即ち、不具合密度と品質問題の有無に関連があると判定できない。
  - － 不具合検出率で同様の分析を行う。

(この際、データの分布が片側に偏っていたため、「対数変換」という手法で偏りを補正する)

→ 結果はP値が0.002となり5%より小さくなった。

よって、不具合検出率が小さい方がリリース後の品質問題が発生する可能性が低い、と判断できる。

— 結論：品質指標としては、不具合密度より、不具合検出率の方が有効である。

#### 事例2：信頼度成長曲線を活用したテスト計画と品質管理

・信頼度成長曲線とは、オープンチャート(時間軸に対する累積不具合数)を数学的なモデルで表現したもの。

$F(t) = K(1 - e^{-\lambda t})$  t:時間、k:総不具合数、 $\lambda$ :不具合の検出能力

— 不具合検出率は、信頼度成長曲線の傾きに相当するので、信頼度成長曲線の式を微分することでモデル化できる。不具合検出率 =  $F'(t) = K\lambda e^{-\lambda t}$

— また、残存不具合件数 =  $Ke^{-\lambda t}$ となる。

— 不具合検出率と残存不具合件数の違いは $\lambda$ 、即ち、不具合の検出能力=テストのやり方である。

つまり、テストのやり方がさほど変わらなければ、残存不具合数は不具合検出率で代替可能である。

・信頼度成長曲線の課題

— テスト開始序盤では、kや $\lambda$ のパラメータ推定値が妥当なものとならず、また終盤においても新たなデータが追加される度にパラメータ推定値は変わってしまう。

— その対策としては、過去の類似プロジェクトのデータを活用してパラメータを推定するという手段がある。

#### <実施手順例>

1. パラメータK(総不具合数)は過去プロジェクトのデータの回帰分析を行って推定する。

2. パラメータ $\lambda$ (不具合検出能力)は過去プロジェクトのデータで $\lambda$ を推定する。

3. 目標不具合検出率の設定は、過去プロジェクトでリリース後の品質が良かったものの最終不具合指摘率の平均から設定する。

4. ここまでの準備ができたなら、モデルを使ってオープンチャートと不具合検出率のグラフを作成し、リリース時点で不具合検出率が目標に収まるか検証する。収まらない場合はテストリソースを調整して目標に収まるよう、テスト計画をシミュレーションする。

#### ■魅力品質のマネジメント

・品質部門が、製品の不具合検出や管理などの「当たり前品質」ばかり追いかけているのは競争力が低下する。

ユーザ視点を持って客観的に製品の良さを評価できる能力が必要である。

→ そのためには統計手法が不可欠となる。

・魅力品質マネジメントプロセス

1. 魅力品質目標を設定・・・“実現可能なレベルの数値目標”ではなく、お客視点で、具体的な比較対象(ライバル社製品)に対して、どれだけ魅力があるか、を具体的に設定。

2. 設計検証・・・納得できる被験者を集め、ユーザ視点での妥当性確認

・魅力品質マネジメントの事例

##### 事例1：音の官能評価

できるだけ客観性を持たせるよう、3つ以上の製品を比較する時でも必ず2つずつの組にして比較する「一対比較法」を用いて、「対象製品」「対象製品の前機種」「競合他社製品」の3つを比較評価した。

##### 事例2：操作性(ユーザビリティ)評価

製品のユースケースを「タスク」としていくつか抽出して被験者に実施してもらい、「有効性(タスクをマニュアル見ずに実施できたか)」「効率性(タスク所要時間)」「満足度(タスク終了後にアンケート)」の3つの観点で評価した。

##### 事例3：魅力品質簡易評価シート

事例1、2は被験者集めなど非常に手間がかかる方法であるため、品質部門メンバーだけで簡易的に評価

を実施できるような評価シートを作成して評価した。

#### 事例4：市場アンケート評価

製品出荷後のユーザアンケート調査を分析した。満足度と重要度の2つの指標で散布図を描き、各アンケート項目がどの象限に位置づけられるかで対策を検討しやすくした「CS グラフ」という分析ツールを導入している。

#### ■本日講演した講義でのノウハウ(コツ)のまとめ

- ・「2群の母平均の差のt検定」は適用範囲の広い万能の統計手法なので、是非、マスターして欲しい。
- ・ソフトウェアメトリクスはレンジの広いデータを扱うことが多い。その偏りを補正するためには「対数変換」が有効である。
- ・ソフトウェアメトリクスは測定誤差が多いので、誤差を少なくする手法を検討すべきである。
  - 個々のデータではなく、「データ群」としてとらえる
  - 順位尺度を用いる
- ・統計ソフトを利用するだけでなく、その原理を理解して応用範囲を広げる。

#### ■質疑応答

Q：音の官能評価、操作性評価を行った際の被験者の数はどのくらいか？

A：経験から8から10人。2群で評価するため、偶数の人数としている。

Q：信頼度成長曲線の事例に関して、貴社では事例を収集できるノウハウを積むまで、どの位の期間を要したか？

A：自社の場合、およそ5年程度要した。

Q：不具合検出率は信頼度成長曲線との傾きに相当する、ということで“瞬間”の値であると理解したが、実際にはある程度の期間で把握する必要があるのではないか？

A：確かに、瞬間の値では変動が大きいため、実際データを取るときは、ある期間での移動平均でデータを取っている。

以上