統計処理はデータ解析が1割！残りは？

～「実践データ・ハンドリング 実務者のためのS言語入門」出版講演～

INABA Consulting office

L Data Science, Co., Ltd.

稲葉 弥一郎

何故、データ・ハンドリングの本を出したのか？

* 医薬統計解析の分野での友人からS言語のデータ・ハンドリングをまとめた本はどこにあるのかを尋ねられた
  * 本屋で「S言語」関係（Rも含む）の本を調べた
    * データ・ハンドリングに関してまとめてある本は無かった
  * 後輩用にまとめたマニュアルがあったのでこれを綺麗にすれば本に出来るのでは？
    * 運良くすべてテキストファイルで保存していた

今まで色々なパッケージを使って感じたこと

* データ・ハンドリングを詳しく書いてあるソフトは無かった
* 仕事で使った解析パッケージ
  * SPSS・Analyst・SAS・S－PLUS
  * VisualStat(英名STATISTICA)
* 皆様からは色々と反論は有るかもしれませんが、私にとって満足のいくマニュアル・セミナーは全く有りませんでした

S言語とS-PLUSとR

* S-PLUS の特徴の一つとして、メニューからマウス操作でデータハンドリングが実行可能
* S言語を習得すれば、より詳細な設定や定型処理化を行う時に便利
  * 解析の再現性が担保される

* R につきましては、S言語と非常に互換性が高く、コマンドをそのままコピー＆ペーストで実行していただけることがほとんどです。
* S言語 及び R の既存のプログラムを資産としてご活用が可能です。

私のバックグラウンド

* 最初はコンピュータメーカーのSE
  * 使う言語は「COBOL・FORTRAN・Assembler」
* システムハウスのSE
* 製薬メーカーのSE
  * 勘定系のシステム設計・開発・運用
    * この時代はパッケージなど無くすべて手作り（COBOLを使ってのシステム開発）
  * 研究所・工場から依頼されたデータの解析
    * FORTRANを使ってのシステム開発
    * PCを使ってのシステム開発
    * データ入力・解析・グラフ出力
    * 自動分析機器のデータ取り込み
  * 研究所での臨床試験のサポート
    * DMシステムの開発（汎用コンピュータ、PC）
    * 統計解析処理（汎用コンピュータ、PC、UNIX）
    * SPSS、Analyst、SAS、S－PLUS
* CROのデータサイエンス部門・システム部門
  * Server・Network整備、DMサポート、統計解析

仕事や独学で覚えた言語・スクリプト

* 言語
  * COBOL・FORTRAN・Assembler
  * PL-1・APL・LISP・Prolog
  * ALGOL・PASCAL・BASIC・C言語
* スクリプト
  * PERL・RUBY・SED・AWK
* DB
  * 富士通のDB各種、IBMのDB各種

統計解析部門の仕事

* 統計部門の業務一覧

  * 統計解析実施前
  * プログラム仕様書・プログラミング・テスト
  * 統計解析実施
  * 統計解析実施後

  * 詳細な説明は次ページ以降

統計解析実施前

* 症例数設計（例数を設定した根拠の説明も含む）
* 薬剤割付設計（治験方法に合わせた割付方法）
* 治験実施計画書の統計部分の作成
* 統計解析手順書(SOP)作成
* 統計解析計画書(SAP)作成
  * 単位・桁数、帳票タイトル等、変更が多いものは、別紙として作成
* 統計解析図表計画書作成
  * CAP: Chart Analysis Planともいう
* 解析実施環境設定書作成

仕様書・プログラミング・テスト

* 統計解析プログラム実施フローの作成
* 統計解析バリデーション計画書作成
  * 統計解析バリデーション計画書、VaValidation確認表
* Validation(SingleまたはDouble)
* 統計解析用DB作成
  * 統計解析用DB仕様書の作成
  * 統計解析用DB作成プログラム仕様書作成
* 統計解析・図表
  * 統計解析プログラム仕様書作成
  * 図表出力プログラム仕様書作成
* 統計Programing(SingleまたはDouble)
* プログラムテスト

統計解析実施

* 統計解析プログラム実施フローに基づく統計解析の実施
* 統計解析報告書作成
* 統計解析バリデーション報告書作成

統計解析実施後

* 成果物まとめ
  * 電子媒体で作成 (実行log含む)

* 総括報告書(CSR)の統計部分(用語・結果)の確認・コメント

初めてS言語を使ったときに感じたこと

* 統計手法とS言語の例題は多くの本が出版されている（インターネット上も含めて）
* データ・ハンドリングの例題集がどこにもなかった（インターネット上も含めて）
* セミナーに参加したが実務に使えるデータ・ハンドリングの教育は無かった

* しかたなくメーカーとQ＆Aを行い実務に使う例題集をまとめていた（標準化を考えながら）
* 社内のマニュアルとして統計実務をしながらまとめていた（標準化を考えながら）

データ・ハンドリングにこだわった理由

* 解析プログラムの9割はデータ・ハンドリング
  * プログラミングとして一番時間が掛かる部分
* 残りの1割が統計解析処理

* 次のページ以降で上記の説明をします

解析プログラム全体の流れ

1. データ入力システムのDBからデータを読込、パッケージのDBに
   変換する
   （複数のテーブルがある（テーブル数として、50程度はある））

2. 「1」で作成したテーブルを組み合わせて、解析しやすいテーブルを作る
   （入力は複数テーブルで出力も複数のテーブルになる）
   （解析結果の帳票形式が多いほどテーブル数も増える傾向にある）

3. 「2」で作成したテーブルを組み合わせて、各種解析を行い、出力テーブルを作る

4. 「3」で作成したテーブルを使い、解析結果を作る
   （テキスト、リッチテキスト、Excel等）

1・2・4はデータ・ハンドリングである、つまり9割がデータ・ハンドリングとなる

本の目次の確認

* 本の目次を確認すると、データ・ハンドリングを主体にまとめてある
* 以下のホームページ参照

* ・サイエンティスト社 新刊案内：
* http://www.scientist-press.com/11_319.html

* ・S-PLUS ホームページ 参考書籍・文献：
* http://www.msi.co.jp/splus/tips/books/newbook19.html

マニュアル化をしようと考えた理由

* 人はすぐに忘れるのでまとめておく必要がある
  * 特にわたしは自分のした仕事も数ヶ月で忘れる
  * 自分でまとめたものは、後で読むと思い出すことが出来る

* マニュアル化は標準化につながる

* 標準化することで仕事の効率が上がる
* 標準化されている道具は使いやすい
* 道具の組み合わせは自由である

* 但し厳然として、組み合わせの標準はある
  * 皆さんはすでに考え方はお持ちでは

私の仕事に対する考え方

* 「早く・正確に・正しい手抜きをする」を目標に仕事をする
* あらゆる道具を使いこなせるように努力する
* 知識は、知っているつもりでも、知らないことのほうが多いので、一生涯学ぶ気持ちを持ち続ける

今後の方向

* R対応をしていけるように努力します
* 今後とも、数理システム様には色々とご協力をお願い致します

ご清聴有難う御座いました