教育・言語データ処理のための「R」活用法

新實 葉子(名古屋大学), 阪上 辰也（広島大学）

USING “R” FOR DATA ANALYSIS IN LANGUAGE TEACHING

Yoko Niimi (Nagoya University), Tatsuya Sakaue (Hiroshima University)

概要: 本ワークショップでは, 実習を通じて統計解析環境の「R」の基本的な利用法を学ぶ。分析対象とするデータは，成績といった教育データだけでなく，コーパス等の言語データの基本的な処理方法も扱う。ワークショップは 5 部構成となり，第 1 部で R とは何かを概説する。第 2 部では R を利用するにあたり重要となる関数・変数の意味と働きを説明する。第 3 部で R による基本的な作図方法を解説し，第 4 部では R で統計的検定を行う方法を説明する。第 5 部では，パッケージ利用による日本語データ処理の方法を紹介する。

【キーワード】統計解析環境 R，関数，変数，データ分析，統計的検定

1. R とは何か

R とは，統計解析処理を行うためソフトウェアである。R の特徴として，「統計処理を行えるソフトウェア」であり，R を利用する主な利点として，1）無償で利用可能，2）統計処理の種類が豊富，3）高度なグラフィックス機能，の 3 点を挙げることができる。

R は，配布サイト（http://www.r-project.org）から，使用 OS に応じたファイルをダウンロードして一般的なソフトウェアと同様のインストール作業を行うことで利用できる。

2. 関数と変数

R の利用には，「関数」と「変数」についての理解が不可欠である。簡潔な定義として，関数は「指定した値に対して何らかの処理をして結果を返すもの」であり，変数は 「2 つ以上の何らかの値（数値や文字列）が格納されたもの」と言うことができる。例えば，R には，平方根を求めるために sqrt という関数が用意されており，括弧内に 144 という数値を与えて関数を実行すると， 12 という結果を返ってくる。この時，単一の値ではなく，関数が処理するものとして，変数を指定すると，複数の値を一括で処理することができる。

3. R による作図方法

R による作図手順は，基本的に，変数に値を代入し，作図用の関数により処理するという 2 段階となる。具体的に作図できるグラフとして，hist 関数を用いたヒストグラム，plot

関数を用いた散布図，boxplot 関数を用いた箱髭図などがある。一例として図 1 に R で出力した箱髭図を示す。

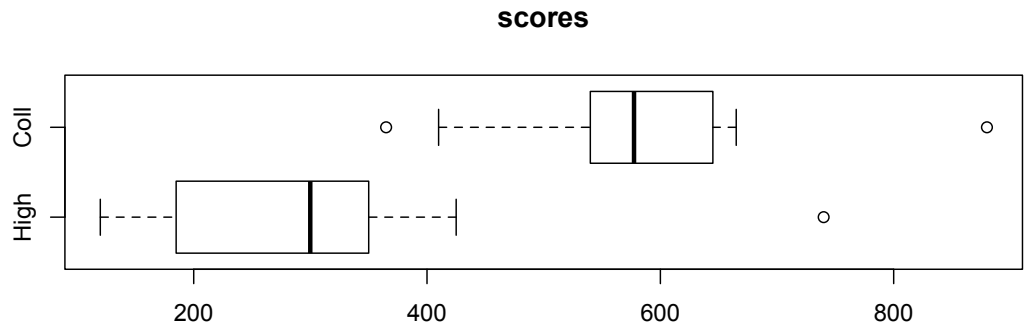

図 1: 箱髭図

4. 統計的検定の実行

言語研究，また教育に関する研究において，統計的検定が必要とされる場面が増えているが，本ワークショップでは， 2 群の平均値の差を検定するために使う t 検定，独立性の検定のために行う χ 二乗検定について，それぞれ t.test と chisq.test という関数を用いて検定を行うことができることを示す。

5. パッケージの利用

最後に，「パッケージ」と呼ばれる R の付加機能を紹介する。基本パッケージだけでも 1000 を超えるパッケージがあり，特殊な処理を行うことができる。本ワークショップでは，言語処理に特化した RMeCab を取り上げ，日本語コーパスの処理において有益なツールであることを示す。

6. おわりに

統計解析ソフトウェアの R は，無償で誰もが利用可能であり，教育・言語データの処理において強力なツールとなる。今回のワークショップを通じて，R の利用者が増え，データ分析に関する情報がさらに共有されることを期待したい。

参考文献

石田基広, R によるテキストマイニング入門, 2008

山田剛史・杉澤武俊・村井潤一郎, R によるやさしい統計学, 2008