外国語教育メディア学会 (LET) 関西支部 メソドロジー研究部会 2011 年度報告論集
小林 雄一郎 (pp. 81–91)

Rによる成績データ分析入門

小林雄一郎
日本学術振興会

概要

本稿の目的は，統計処理環境 R を用いた成績データ処理の基礎を紹介することである。具体的には，数十クラス，数百人の成績データからクラスごとの傾向，学部ごとの傾向，男女ごとの傾向，教員ごとの傾向といった有益な情報を抽出して視覚化し，統計処理を行う方法を説明する。

Keywords: 統計処理環境 R，成績データ，統計処理，視覚化

1. はじめに

教育現場には多くの成績データが存在する。では，大量の成績データを効率的に処理し，そこから有益な情報を見つけ出すにはどうしたらいいのか。

本稿では，統計処理環境Rを用いて，数十クラス，数百人の成績データからクラスごとの傾向，学部ごとの傾向，男女ごとの傾向，教員ごとの傾向といった有益な情報を抽出して視覚化し，統計処理を行う方法を説明する。なお，主にWindowsでの操作を想定しているが，必要に応じて，Macでの処理方法にも言及する。

2. Rとは何か

R とは，統計処理とグラフィックスのためのフリーの統計解析環境である。これは，1991 年に Ross Ihaka 氏と Robert Gentleman 氏によって開発が始められ，1993 年 8 月に産声を上げた。R の主な利点としては，(1) フリーソフトであるため，誰でも無料で利用することができる，(2) Windows，Mac，Linux など，様々な OS 上で動作させることができる，(3) グラフィックスの機能が非常に充実している，(4) 拡張機能が「パッケージ」という形で配布されており，それらを誰でも無料でダウンロードできるため（http://www.r-project.org/），最新の統計解析手法をすぐに試すことができる，などである。

R に関する情報のかなりの部分は，インターネット上で得ることができる。また，R に関する書籍は，日本語でも数多く出版されている。R のダウンロードやインストールから丁寧に解説した書籍としては，舟尾・高浪 (2005) やジュール他 (2010) などが分かりやすい。

3. Rの基本的操作

3.1 コマンド入力

R を起動するには，スタートメニューのプログラム，あるいはデスクトップにあるアイコンをクリックする。すると，図 1 のようなコンソール画面が現れる（R のバージョンによって若干見た目が異なる場合があるが，実際の操作に関して大きな違いはない)。

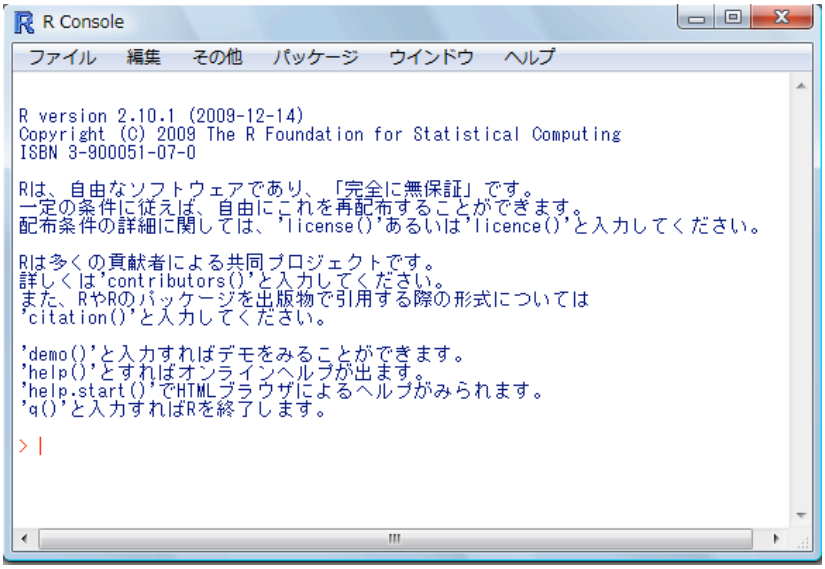

図 1. R のコンソール画面 (R version 2.10.1)

このコンソール画面に「コマンド」と呼ばれる命令を入力すると，その答えが返ってくる。例えば，`(1+2)*(3-4)/5` という命令を入力すると，`-0.6` という結果が得られる。

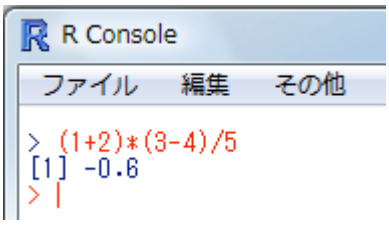

図 2. R のコマンド入力

このように，R による処理は，コマンド（赤字）を入力すると，結果（青字）が返ってくるという形式となっている。

3.2 記述統計

R にデータを読み込む方法はいくつかあるが，少数のデータを手作業で入力する場合

は，c 関数を用いる。図 3 は，5 人分の英語の成績を Eng という変数に代入したあと，変数 Eng の中身を表示した例である。

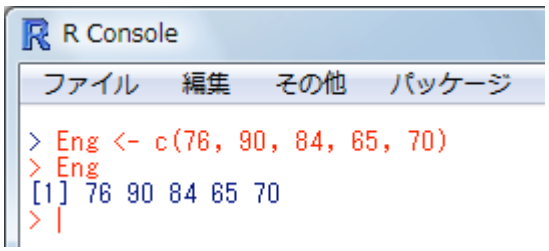

図 3.　c 関数によるデータ入力

なお，ここでいう「関数」とは，R で何かを実行するための命令のことであり，「変数」とは，何らかの処理結果を一時的に入れておくための箱のようなものである。

記述統計に関する関数としては，データの総和を求める sum 関数，平均を求める mean 関数，最大値を求める max 関数，最小値を求める min 関数，標準偏差を求める sd 関数，不偏分散を求める var 関数などがある。また，summary 関数を用いると，データの最小値，下側 25%点，中央値（50%点），平均値，上側 25%点，最大値が一度に表示される。図 4 は，これらの関数を変数 Eng のデータに対して適用した結果である。

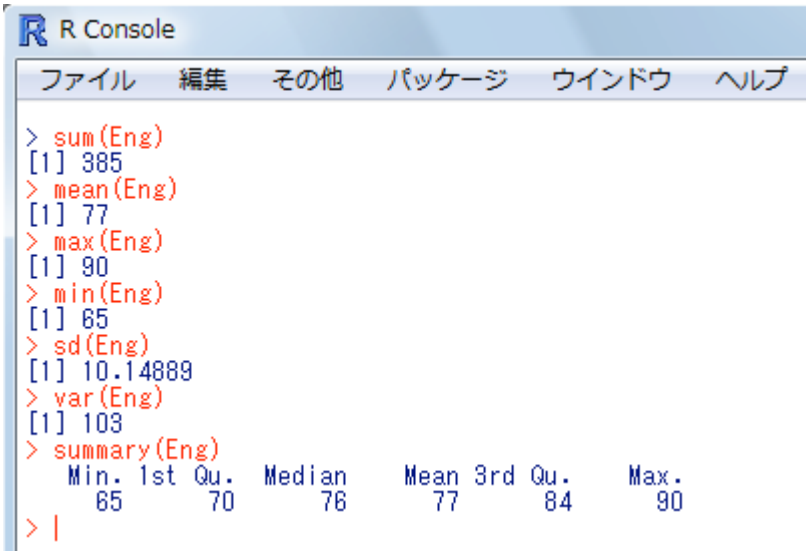

図 4.　記述統計

4.　大規模な成績データの処理

4.1　Excel データの読み込み

前節では手作業でデータを入力したが，実際のデータは数十人，数百人，場合によっては数千人に及ぶこともある。そこで，この節では，Excel ファイルに保存されたデータ

を R に読み込む方法を説明する。表 1 のような 800 人分の成績データがあったとする。表中の student，class，sex，faculty，score は，それぞれ学生 ID，クラス，性別，学部，点数を表している。また，クラスには 1〜20 組の 20 クラス，学部には A〜J 学部の 10 学部がある。

表 1

800 人分の成績データ（一部）

student	class	sex	faculty	score
1	1	M	A	72
2	1	F	A	94
3	1	M	A	90
4	1	F	A	88
5	1	M	A	70
…	…	…	…	…
800	20	M	J	48

Windows における読み込みの手順としては，(1) Excel ファイルを開いて，分析対象となる表の全体（ヘッダー部分も含む）を `Ctrl＋c` などでコピー，(2) R のコンソールで `read.delim` 関数を使ってクリップボードからデータを読み込んで，変数に代入，という 2 つのステップがある。また，Mac では，`read.delim("clipboard")` の代わりに，`read.delim(pipe("pbpaste"))` を用いる。図 5 では，表 1 のデータを読み込んで，`dat001` という変数に代入している。

図 5.　クリップボードからのデータ読み込み (Windows)

4.2　ヒストグラム

読み込んだデータにおける点数（＝データの 5 列目）の概要を見るには，前出の `summary` 関数を用いる。また，ヒストグラムを描いて，データの全体像を視覚化するには `hist` 関数を用いる。なお，`hist` 関数には，「引数」と呼ばれるオプションが存在し，色を指定する `col`，X 軸のラベルを指定する `xlab`，Y 軸のラベルを指定する `ylab`，表のタイトルを指定する `main`，X 軸の目盛の下限と上限を指定する `xlim` などがある。

```
R Console
ファイル　編集　その他　パッケージ　ウインドウ　ヘルプ
> summary(dat001[, 5])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.00   54.00   65.00   64.06   73.00   98.00
> hist(dat001[, 5], col="red", xlab="score", ylab="student", main="Test", xlim=c(0, 100))
> |
```

図 6.　データの要約とヒストグラムの描画

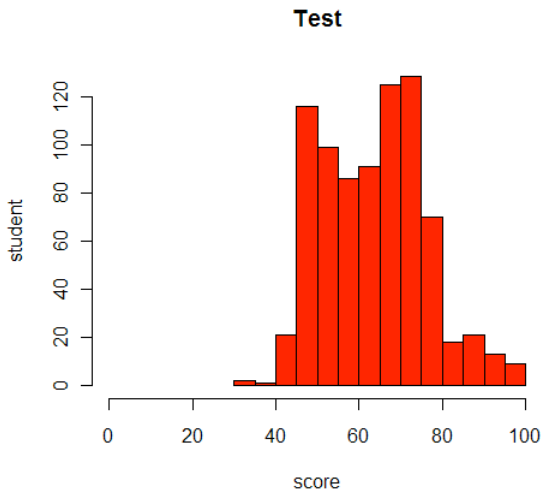

図 7.　[図 6]の実行結果として得られるヒストグラム

4.3　箱ひげ図

次に，箱ひげ図を描くことで，クラス間の差を可視化する。まず，split 関数を用いて，変数 dat001 の中のデータを集計する。図 8 は，dat001 の中の点数（＝データの 5 列目）を，クラス（＝データの 2 列目）別に集計した結果である。

```
R Console
ファイル　編集　その他　パッケージ　ウインドウ　ヘルプ
> class.dat <- split(dat001[[5]], dat001[[2]])
> class.dat
$`1`
 [1] 72 94 90 88 70 82 86 92 72 90 58 72 82 86 78 94 96 88 84 92 74 52 66 71
[25] 94 92 82 82

$`2`
 [1] 70 94 94 54 92 68 90 68 80 74 92 84 68 86 98 74 82 72 98 82 52 42 84 68
[25] 36 80 86 96 72
```

図 8.　クラス別の集計

なお，split(dat001[[5]], dat001[[3]])とすると男女別の集計となり，split(dat001[[5]], dat001[[4]])とすると学部別の集計となる。

そして，クラス別に集計した結果で箱ひげ図を描くには，boxplot 関数を用いる。この boxplot 関数の引数には，色を指定する col，X 軸のラベルを指定する xlab，Y 軸のラベルを指定する ylab，Y 軸の目盛の下限と上限を指定する ylim などがある。

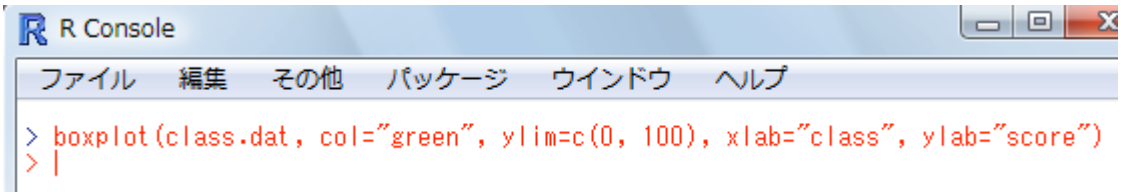

図 9.　箱ひげ図の描画

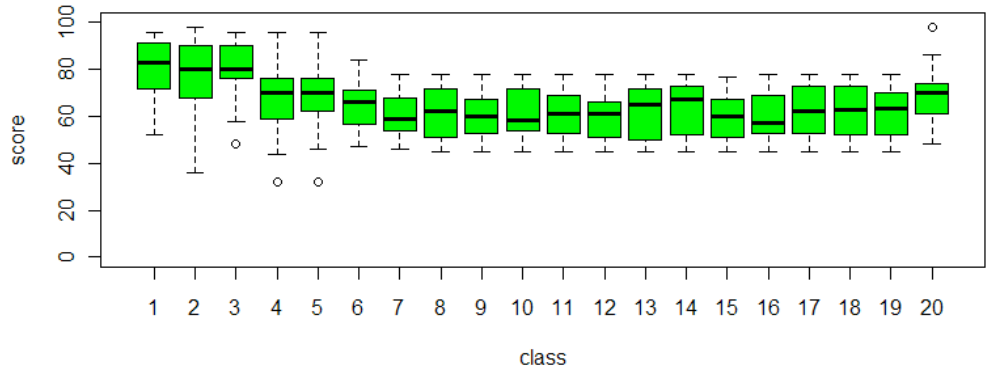

図 10.　[図 9]の実行結果として得られる箱ひげ図

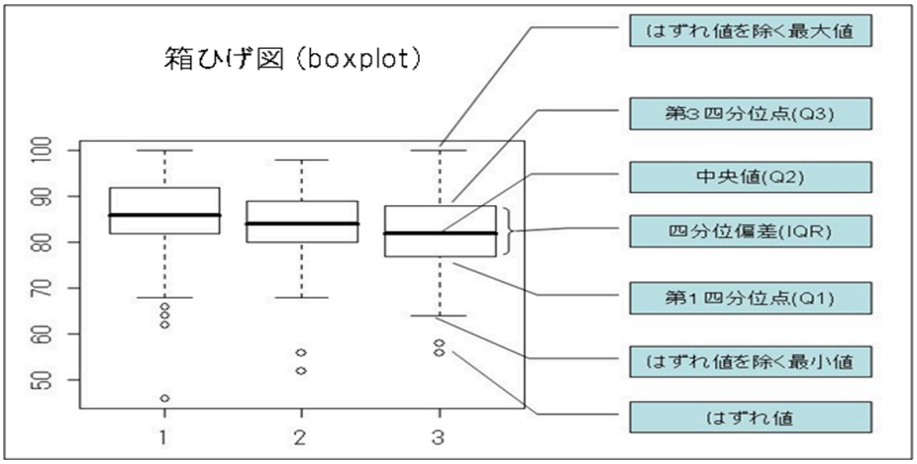

図 11.　箱ひげ図の見方

図 10 を見ると，1〜3 組の成績が高いこと，2 組の成績にバラツキが大きいこと，いくつかの組にはずれ値が見られること，などが分かる。

やや複雑な処理にはなるが，図 10 を中央値の低い順に並びかえるには，図 12 のような処理を行う（コマンドが長いため，2 行に分けて入力している）。

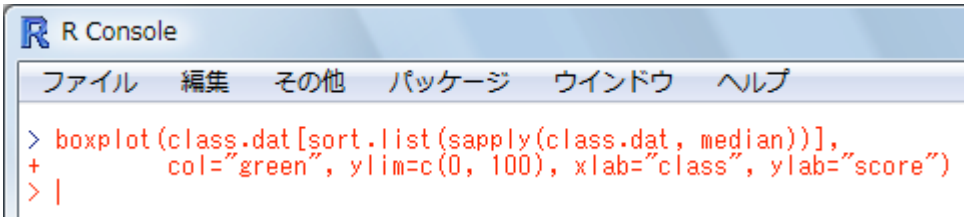

図 12.　箱ひげ図の並びかえ

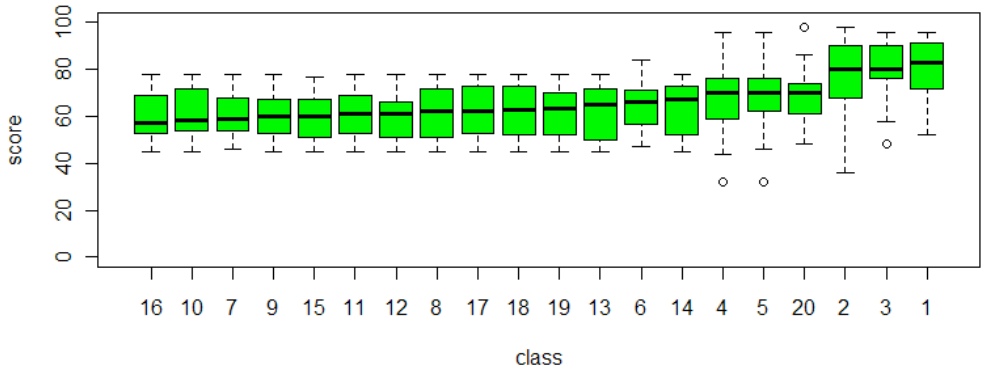

図 13.　[図 12]の実行結果として得られる箱ひげ図

4.4　クラス別の記述統計

クラス別に集計したデータに sapply 関数を適用すると，クラス別の記述統計を簡単に得ることができる。

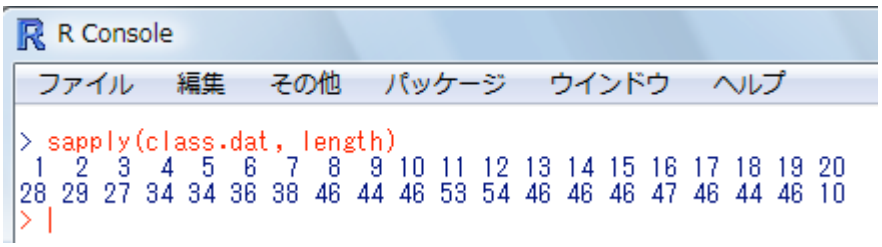

図 14.　クラス別の人数

図 14 では，sapply 関数を使って，クラス別の length（データの長さ＝人数）を求めていて，これを見ると，1 組の人数が 28 人で 2 組の人数が 29 人であること，などが分かる。なお，length 以外にも，mean，max，min，sd，var といった様々な記述統計量を求めることができる。

5. テスト結果の比較

5.1 散布図

この節では，中間テストと期末テストの比較のように，2 つのテスト結果を比較する手法を見ていく。

表 2 のような 50 人分の成績データがあったとする。表中の Student，Test A，Test B は，それぞれ学生 ID，中間テストの点数，期末テストの点数を表している。

表 2
中間テストと期末テストの結果（一部）

Student	Test A	Test B
1	51	58
2	45	27
3	75	83
4	60	60
5	41	37
…	…	…
50	72	54

まず，前出の read.delim("clipboard") を用いて，表 2 のデータを変数 dat002 に代入する。そして，2 列目のデータ（中間テストの結果）を変数 test.A に代入し，3 列目のデータ（期末テストの結果）を変数 test.B に代入する。

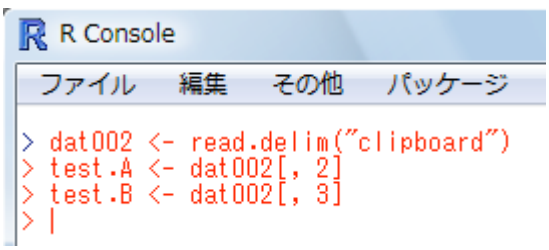

図 15. テスト結果の読み込み

読み込んだデータを使って散布図を描くには，plot 関数を用いる。なお，plot 関数の引数には，X 軸のラベルを指定する xlab，Y 軸のラベルを指定する ylab，表のタイトルを指定する main，X 軸の目盛の下限と上限を指定する xlim，Y 軸の目盛の下限と上限を指定する ylim などがある。そして，図 16 では，図中に直線を引く関数である abline 関数を用いて，それぞれのテストの平均点の値に点線を描いている（引数 v は vertical，h は horizontal，lty は line type をそれぞれ表す）。また，test.A と test.B で回帰分析を行うには，lm 関数を用いる。そして，この lm 関数と abline 関数を組み合わせることで，散布図上に回帰直線を引くことができる。

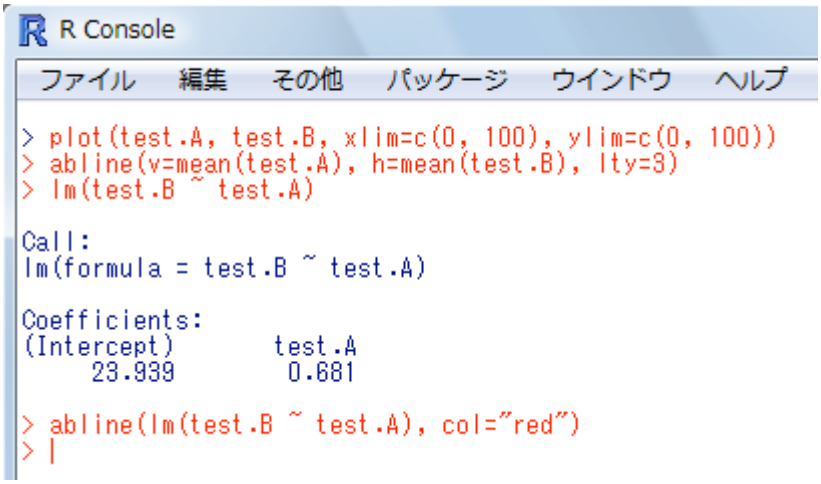

図 16.　散布図と回帰直線の描画

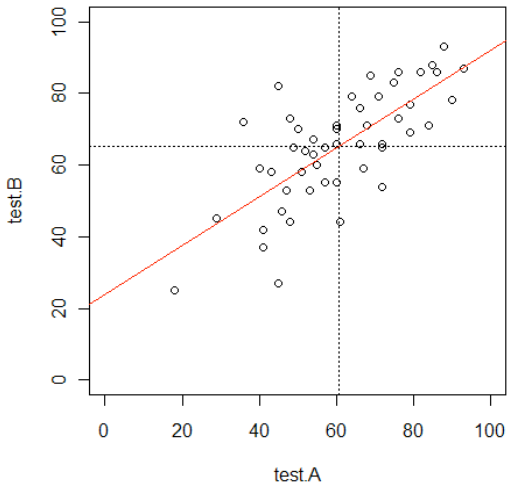

図 17.　[図 16]の実行結果として得られる散布図

5.2　相関係数

中間テストの結果と期末テストの結果の相関係数を求めるには，cor 関数を用いる。その際，引数 method で pearson を指定するとピアソンの積率相関係数を返し，spearman を指定するとスピアマンの順位相関係数を返す。

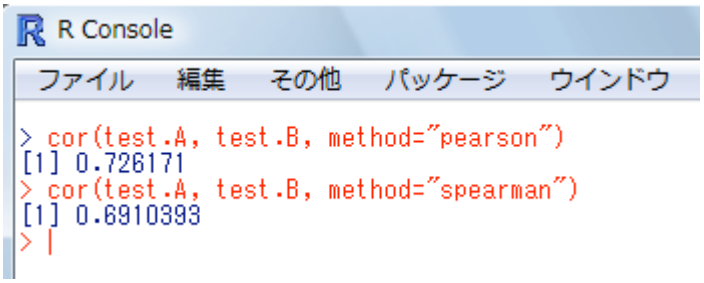

図 17.　相関係数の計算

5.3　*t* 検定

中間テストの結果と期末テストの結果に *t* 検定を行うには，t.test 関数を用いる。引数 paired で TRUE を指定すると対応ありの *t* 検定となり，FALSE を指定すると対応なしの *t* 検定となる。また，引数 var.equal で TRUE を指定すると等分散性を仮定し，FALSE を指定すると等分散性を仮定しない（Welch の方法）。

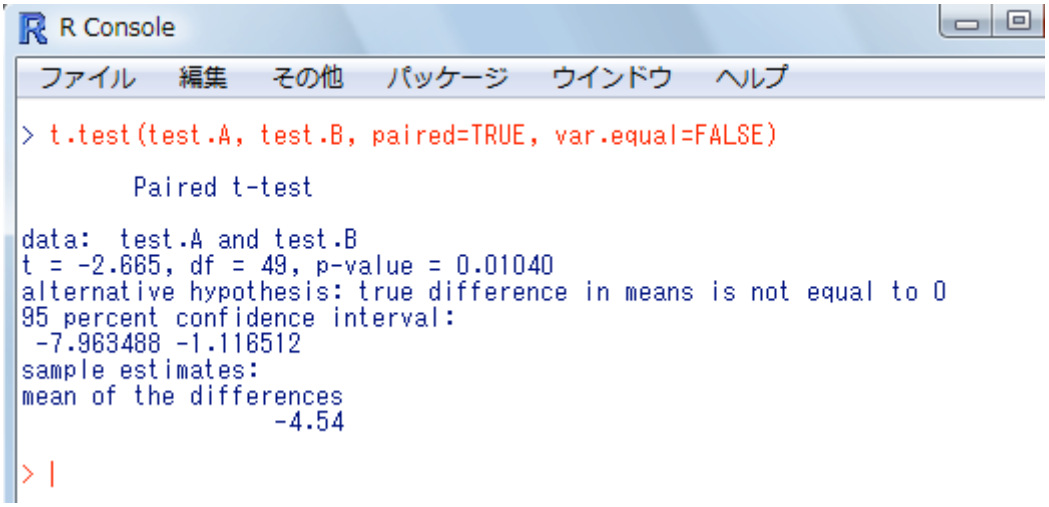

図 18.　対応ありの *t* 検定（Welch の方法）

図 18 を見ると，*t* 検定の結果として得られた p-value は 0.01040 であり，中間テストの結果と期末テストの結果の間には，5%水準で統計的に有意な差があることが分かる。

6. おわりに

本稿では，R を用いた成績データ処理の基礎を解説してきた。R の機能は多種多様にわたり，紙面の都合で割愛した内容も多い。そこで最後に，R を学ぶ言語教育研究者に向けて，参考文献を紹介する。

まず，R のインストールや基本操作に関しては，前述のように，舟尾・高浪 (2005) とジュール他 (2010) が非常に分かりやすい。また，R による統計処理に関しては，山田他 (2008)，青木 (2009)，服部 (2011) などがおすすめである。そして，R 関連ではないが，教育データの統計処理に関する解説書としては，三浦 (2004) が白眉である。

謝辞

本稿の内容は，2011 年 8 月 6 日（土）に名古屋学院大学で行われた外国語教育メディア学会 (LET) 2011 年度大会のワークショップ「R による教育データ分析入門」の一部に基づいている。ワークショップの機会を与えてくださった尾関修治先生（名古屋大学）と阪上辰也先生（広島大学），そして，本稿執筆の機会を与えてくださった水本篤先生（関西大学）に心より御礼を申し上げます。

参考文献

青木繁伸 (2009). 『R による統計解析』オーム社.

舟尾暢男・高浪洋平 (2005). 『データ解析環境「R」』工学社.

服部環 (2011). 『心理・教育のための R によるデータ解析』福村出版.

三浦省五 (編) (2004). 『英語教師のための教育データ分析入門―授業が変わるテスト・評価・研究』大修館書店.

山田剛史・杉澤武俊・村井潤一郎 (2008). 『R によるやさしい統計学』オーム社.

A. ジュール・E. イエノウ・E. ミーターズ (2010). 『R 初心者のための ABC』シュプリンガー・ジャパン.