研究レポート

データ解析環境Ｒによる統計的品質管理

荒　木　孝　治*

１．はじめに

品質管理のみならずあらゆる分野において、データ解析時にコンピュータの利用は欠かせない。そのとき、データ解析用ソフトウェアの選択肢には商用のものとフリーのものがある。代表的な商用ソフトとして、JUSE-StatWorks（http://www/i-juse.co.jp）やS-Plus（http://www.msi.co.jp）、JMP（http://www.jmp.com）、SPSS（http://www.spss.co.jp）などがあるが、これらを企業や学校で大量に導入したり、個人で購入したりするには費用の面から難しい。そのため、品質管理や統計の教育を受けてもそれを実践しないまま知識を錆つかせてしまうという状況が生まれる。データ分析の教育内容は理論的であるとともに高度に実践的でもあるので、自分の関心領域で適用してはじめて身につく。また、表計算ソフトを利用して、非能率的な作業を行っているという状況も多く見受けられる。

これに対してフリーのオープンソースソフトウェアがある。データ解析の分野におけるその最強のものとして、世界の第一線の研究者や実務家が協力しあって開発しているデータ解析環境Ｒを挙げることができる[20]。本稿では、Ｒの品質管理での利用可能性を検討していく。

２．データ解析環境Ｒ

入門的なデータ解析では、マイクロソフト社のエクセルで十分であるという意見がある。しかし、これには注意が必要である。なぜならエクセルについてはアルゴリズムが公開されておらず、内部で行われている計算が“見える化”されていないという問題を持つ。例えば、乱数の信頼性や統計量・分布関数の計算精度など確率・統計計算における様々な問題が報告されている[18, 19][1]。これらより、厳密な結果が要求される場合には、エクセルを利用する

＊商学部　教授

1　演算子の優先順位が特異な形で割り付けられているという問題も指摘されている[17]。例えば、-3^2は（-3）^2と解釈され、9となる（-3^2=9）。

ことに注意が必要となる。また、デフォルトのグラフの貧弱さから、グラフの完成度を高めるために、時間を消耗する傾向がある。あるいは、作成者がグラフの理想的なスタイルを知らないため、貧弱なグラフで良しとする傾向も見られる。信頼のおけるアドインソフトウェアを利用すれば別であるが、エクセルを用いて、“正しい”ヒストグラムや散布図、パレート図を作成するのは大変である。ただし、データのハンドリングに関して、エクセルに代表される表計算ソフトウェアは便利なので、この機能は積極的に利用してよいだろう。データを表計算ソフトウェアで入力し、それをテキスト形式で保存すると、多くのより高度な統計解析ソフトウェアでデータファイルの読み込みが可能となる。

Rは次のような特長を持つソフトウェアである。

- オープンソースである

  ソースコードが公開されている。そのため、日々、内容に対するチェック機能が働いている。また、過去から蓄積されてきた信頼できる計算ライブラリが利用されている。

- フリーである

  個人利用はもとより、企業や学校でフリーに利用できる。また、改変したり機能を追加したりして、一定のルールの下で配布できる。学校や企業における不正コピーの問題から解放される。

- さまざまな分析手法を利用できる

  入門的な手法から高度な手法まで利用できる。また、最新の論文で発表された統計量やモデルが次々とRに実装されている。

- プログラミング言語である

  定型の分析だけでなく、目的に応じて機能を拡張したり、新しい機能を付加したりすることができる。プログラミングの入門として、利用することも可能である。

- 各種OSに対応している

  Windows、Mac OS、Linux、Unixといった主要プラットフォームで利用できる。

Rはこうした利点を持つが、利用の仕方がCUI（Character User Interface）を基本とすること、基本は英語版という問題も一部の利用者に対しては存在する。CUIというのは、キーボードからコマンド（関数）を入力して、計算や作図等の命令を実行する方式で、英語に基づくコマンドを入力する必要があることから、初心者にとっては難しいと感じる主要因の1つとなる。CUIに対して、メニューやアイコンを用いて命令を実行するGUI（Graphical User Interface）という方式が、WindowsやMac OSにおける標準の方法である。英語の問題に関しては、2005年4月にR本体の国際化バージョンがリリースされ、本格的に日本語化が進んだ。その結果、メッセージの一部およびメニューが日本語化されるとともに、日本語データも取り扱うことができるようになった。また、Rを用いたデータ解析に関する本が続々と出版されており[1－16]、日本語によるRの利用環境はかなり整った。さらに、日本

語によるRに関する様々な情報が、RjpWiki（http://www.okada.jp.org/RWiki）に日々蓄積されており、充実した情報源となっている。

Rは基本パッケージと拡張（貢献）パッケージから構成される。基本パッケージで、基本的な分析を実行できる。拡張パッケージを用いると、より高度で適用領域に特化した分析が可能となる。拡張パッケージの数は、2007年3月現在で1000に近い。拡張パッケージは、利用者の必要に応じてインストールする。また、標準の設定では、Rを起動したときに利用可能なのは主に基本パッケージで、インストール済みの拡張パッケージは、必要に応じて利用可能にする必要がある。

Rを起動するには、Rをインストールした際にデスクトップに作成されているショートカットをダブルクリックする。すると図1に示すR Console（Rコンソール）が表示される[2]。

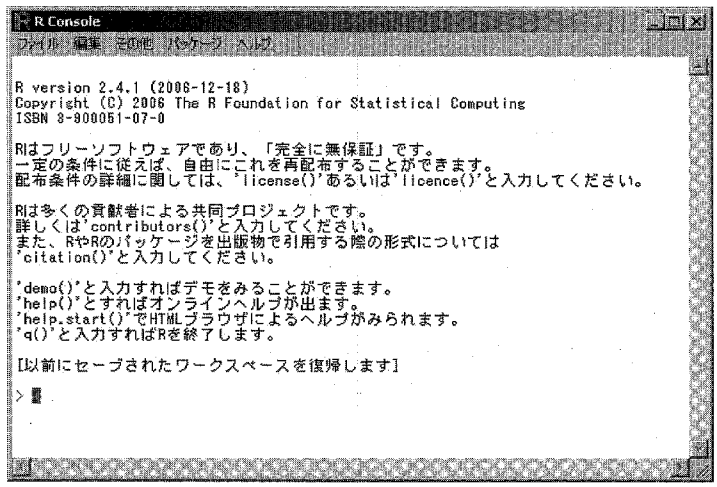


図1　Rコンソール

3．品質管理でのRの利用

統計的品質管理（SQC：Statistical Quality Control）に関連した基本的な手法に関して、Rでできることを表1に示す。表1からわかるように、品質管理における基本的な手法であるQC七つ道具が利用可能である（評価欄の記号は、◎：十分、○：普通、△：なんとか利用できる、ことを意味する）。なお、表中のパッケージqcc（Quality Control Charts）は統計的品質管理用の拡張パッケージである[21]。これを用いると、計量値・計数値データに対するシューハート管理図、Cusum管理図、EWMA管理図、工程能力分析、パレート図、特性要因図等の分析を行うことができる（Rcmdrに関しては後述する）。

2　以下、出力はWindows版のR（バージョンは2.4.1）による。

表1　SQC手法とパッケージ

手法				評価	パッケージ
QC七つ道具	グラフ	折れ線グラフ		◎	Rcmdr
		棒グラフ			
		円グラフ			
		管理図	$\overline{X}-R,\ X-R,\ \overline{X}-s$ $P,\ pn,\ u,\ c$	○	qcc
	パレート図			○	qcc
	ヒストグラム			○	Rcmdr
	特性要因図			△	
	散布図			◎	
	層別			◎	各手法の中で
他	箱ひげ図			◎	Rcmdr
	QQプロット				

QC七つ道具の他に、次に示すより進んだ諸手法を利用できる（記号「*」は、後述の拡張パッケージRcmdrで利用可能な手法をあらわす）。

- １標本、２標本の検定・推定*
- 分散分析・実験計画法*
- 多変量解析・データマイニング

  回帰分析*（単回帰、重回帰、非線形回帰）、一般化線型モデル*、クラスタ分析*、判別分析、主成分分析*、因子分析*、共分散構造分析、決定木、ニューラルネット
- 時系列解析
- ノンパラメトリック法*
- 計算機を集約的に利用する手法

  ブートストラップ法、マルコフ連鎖モンテカルロ（MCMC）法

このようにRには、品質管理の世界で利用されている手法のみならず、まだ十分に取り込まれていない多くの手法も用意されている。これらが１つのソフトウェアで統一的に、しかもフリーで利用できることは大きな利点である。これに対して利用の障壁となる要因として先述のCUIであることが残る。しかし、この欠点を解消する方法がある。その方法の１つが、GUIで利用するための拡張パッケージであるRコマンダー（Rcmdr: R Commander)[3]の利用である。これを用いることにより標準的な手法がメニュー方式で利用可能となる。また、用意されていない手法も独自に拡張することをサポートする機能を備えている。2005

3　カナダのMcMaster大学のJ. Fox教授が開発したGUIパッケージである（http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/)。GUI化のパッケージは他にもあるが、メニュー化された手法の充実性から本稿ではRcmdrに注目した。

年9月に国際化バージョンが公開された[4]。

しかし、まだ問題は残っている。SQCは、特にグラフに関して、長年培われてきた洗練された独自のスタイルを持つ。たとえば、ヒストグラムを作成するとき、データの測定単位を用いて区間の取り方を精妙に調整したり、考察するときに必要な情報（データ数n、平均の位置・値$\bar{x}$、標準偏差s、規格値）を記入したりする。管理図やパレート図といった他のグラフにも独自のスタイルがある。グラフをこうしたスタイルで、ある程度簡単に作成できない場合、RがSQCのプラットフォームとして普及するのは難しい。特に、SQCの基本手法であるQC七つ道具に関しては、できるだけ日本的SQCのスタイルにあった形で実現する必要がある。また、QQプロットや箱ひげ図、密度推定等、SQC以外の世界では標準となっている手法の取り込みも必要である。

Rcmdrには、ある程度簡単に、利用者が独自にGUI機能を付加することができる仕組みが用意されている。例えば、フレームを用意して、その中にラベル、チェックボックス、ラジオボタン、リストボックス、コマンドボタンなどを配置してダイアログボックスを作成することができる。そこで、こうした機能を利用して、SQCの標準的な手法に関してはGUIでの分析を可能とする活動を行ってきた。その成果がQC7toolsであり[5]、これを利用してQC七つ道具および問題解決の考え方・手順をまとめたものが荒木[2]である。

4．RcmdrとQC七つ道具

Rcmdrを起動すると、図2に示すウィンドウが表示され、基本的な機能、手法の利用が可能となる。Rcmdrのウィンドウは、上方よりメニューバー、ツールバー、スクリプトウィンドウ、出力ウィンドウから構成される。QC七つ道具他の機能は、《グラフ》メニューのサブメニューとして組み込んでいる（図3）。

Rcmdrに追加したQC七つ道具（QC7tools）の機能は、2007年3月現在で、折れ線グラフ、棒グラフ、比率グラフ、レーダーチャート、ヒストグラム・工程能力指数（C_p, C_{pk}）、計量値の管理図（$\overline{X}-R$, $X-R$, $\overline{X}-s$）、計数値の管理図（p, pn, u, c）、パレート図である。また、分散分析への拡張を考えて、ドットチャートと交互作用チャートを実装している。これによりQC七つ道具および分散分析の作図を行うことができる。RcmdrとQC七つ道具のグラフの代表的なものを図4から図10に示す。これらの図でキャプションの括弧内の「Rcmdr」は、Rcmdrが本来備える機能であることを示し、「QC7tools」は、Rcmdrに付加したQC七つ道具の機能であることを示す。

4　荒木が日本語訳を提供し、現在、日本語での利用が可能である。他に、スペイン語、ポルトガル語、ルーマニア語、ロシア語等で利用可能となっている。

5　筆者のホームページ（http://www.ec.kansai-u.ac.jp/user/arakit/index.htm）よりダウンロード可能。

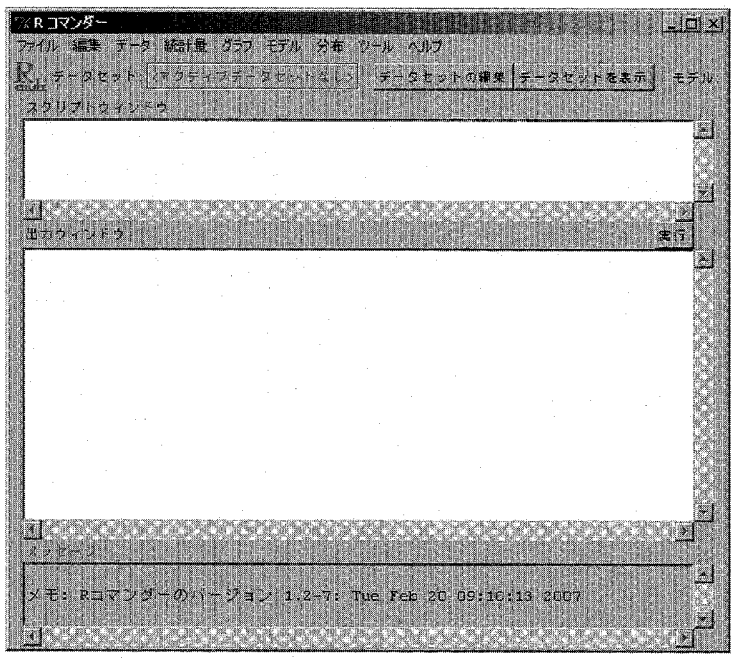


図2　Rcmdrの起動画面

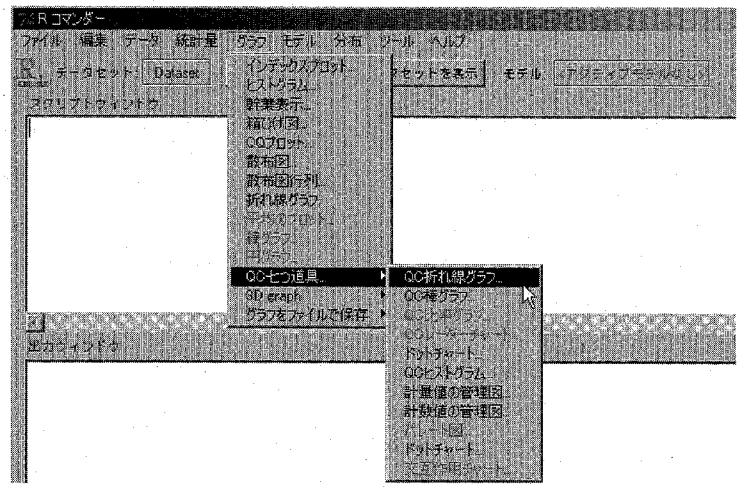


図3　Rcmdrへの追加機能（QC七つ道具）[6]

6　薄いグレイで表示されているメニューは、現状では選択できないことを意味する。Rcmdrでは、手法の誤用をできるだけ避けるため、手法に適合するデータセットが利用可能になっていないと、その手法を利用できなくする工夫を行っている。

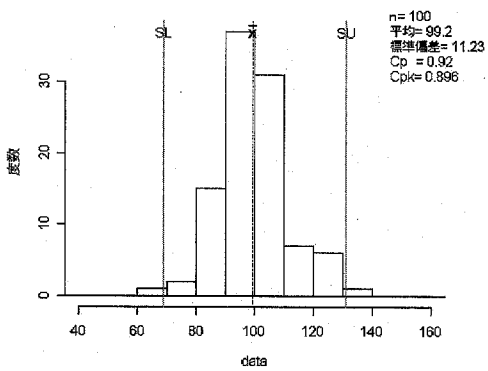


図４　ヒストグラム（QC7tools）

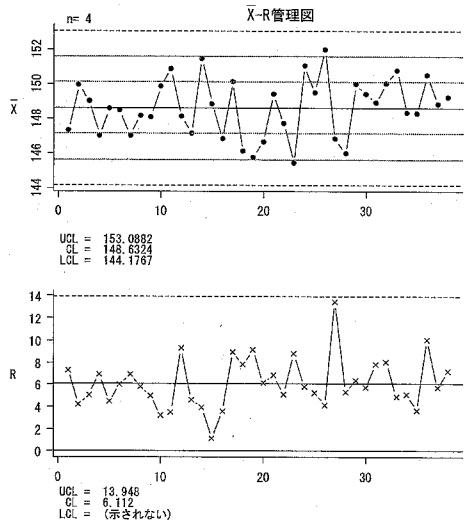


図５　$\overline{X}-R$管理図（QC7tools）

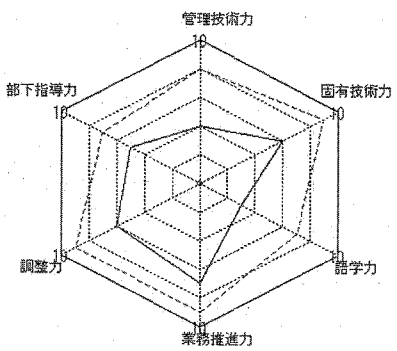


図６　レーダーチャート（QC7tools）

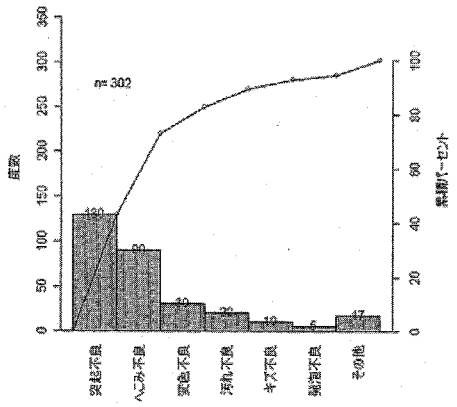


図７　パレート図（QC7tools）

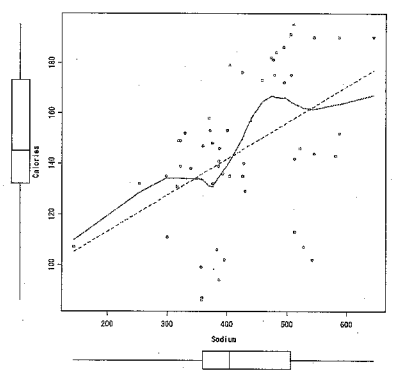


図８　散布図（Rcmdr）

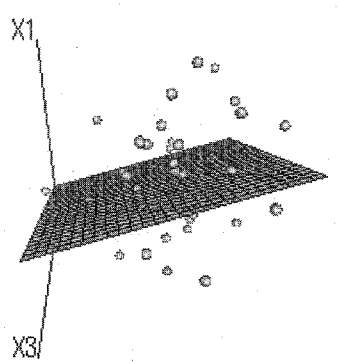


図９　３次元散布図（Rcmdr）

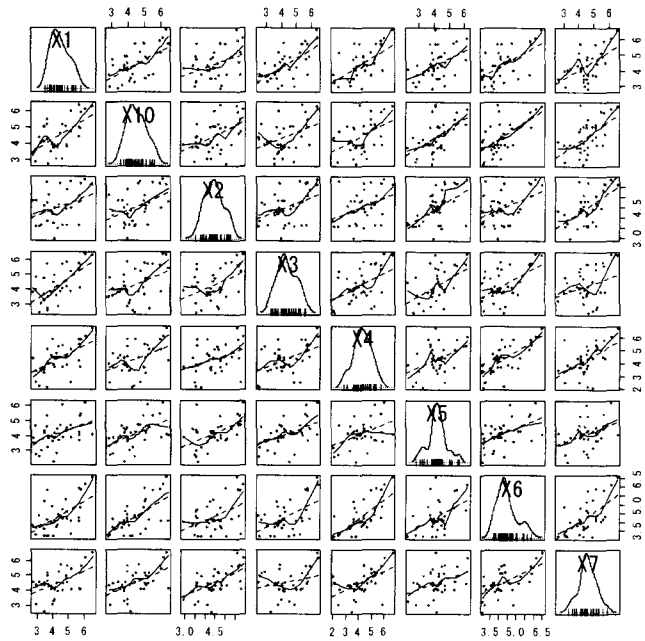


図10　散布図行列（Rcmdr）

5．少し高度な手法— 2元配置分散分析

実験計画法の基本となる2元配置分散分析という少し高度な手法を取り上げ、Rによる解析の手順を見る。2つの因子（A：2水準、B：3水準）を取り上げ、繰り返し2回の完全無作為化実験を行ったところ、表2に示す特性値のデータを得た。このデータに基づいて要因効果について検討したい（値は、大きい方が良いとする）。

表2　データ表

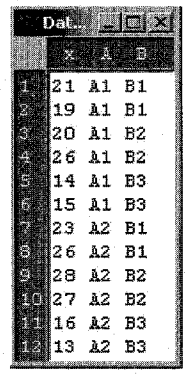
	B_1	B_2	B_3
A_1	21 19	20 26	14 15
A_2	23 26	28 27	16 13

Rcmdrでこのデータを分析するには、図11に示す形式でデータを入力する。変数xにデータ、変数AおよびBに、データに対応する処理記号（A_1, A_2およびB_1, B_2, B_3）を入力している。

Rcmdrの《統計量》メニューの《モデルへの適合》より《線型モデル》を選択し、

x～A＋B＋A：B

というモデルを適用する。これは変数xを、A、B、A：Bの線型式で説明するという線型モデルであり（A：Bは、因子AとBの交互作用A×Bを意味する）、次に示す2元配置分散



	x	A	B
1	21	A1	B1
2	19	A1	B1
3	20	A1	B2
4	26	A1	B2
5	14	A1	B3
6	15	A1	B3
7	23	A2	B1
8	26	A2	B1
9	28	A2	B2
10	27	A2	B2
11	16	A2	B3
12	13	A2	B3

図11　データの表示

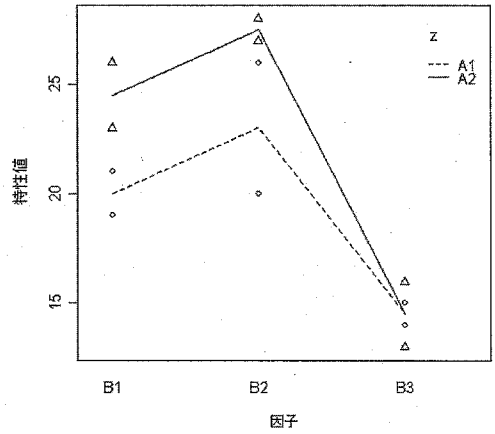


図12　交互作用プロット（QC7tools）

分析のデータの構造式に対応するＲのモデルの記法である。

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

次に、《モデル》メニューの《仮説検定》より《分散分析表》を選択すると、Rcmdrの［出力ウィンドウ］に分散分析表（Anova Table）が表示される（図13）。

出力ウィンドウ

```
> Anova(LinearModel.1)
Anova Table (Type II tests)

Response: x
          Sum Sq Df F value    Pr(>F)
A          27.00  1   5.400 0.059141 .
B         246.17  2  24.617 0.001282 **
A:B        13.50  2   1.350 0.328017
Residuals  30.00  6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

図13　分散分析表

結果は英語で出力されているが、「Sum Sq」は平方和（Sum of Squares)、「Df」は自由度（Degrees of freedom)、「F value」はＦ値、「Pr（＞F)」はＰ値、「residuals」は残差（誤差）であることを理解していれば解釈は容易である。分散分析表より、交互作用Ａ×Ｂは有意ではなく、Ｐ値（0.328）も20％を超えていることがわかる。交互作用を誤差にプールして分散分析表を作成し直すと、図14になる。これより主効果Ｂは高度に有意であり、主効果Ａは有意ではないが無視できない。このモデルに対する残差の正規ＱＱプロット（図15）より、残差の正規性を疑うべき根拠はないと判断できる。

交互作用をプールしたモデルで、因子ＡとＢの水準組合せの母平均を推定すると、図16となり、これをグラフ化すると図17となる。これより、母平均を最大とする水準組合せは、A_2B_2であり、その点推定値は26.4、信頼下限は23.6、信頼上限は29.2である。

以上見てきたように、Rcmdrを用いると２元配置分散分析における一連の作業を、メニューを選択しながら簡単に実行できる。

```
出力ウィンドウ
> Anova(LinearModel.1)
Anova Table (Type II tests)

Response: x
          Sum Sq Df F value    Pr(>F)
A          27.00  1  4.9655 0.0564426 .
B         246.17  2 22.6360 0.0005086 ***
Residuals  43.50  8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

図14　交互作用をプールした分散分析表

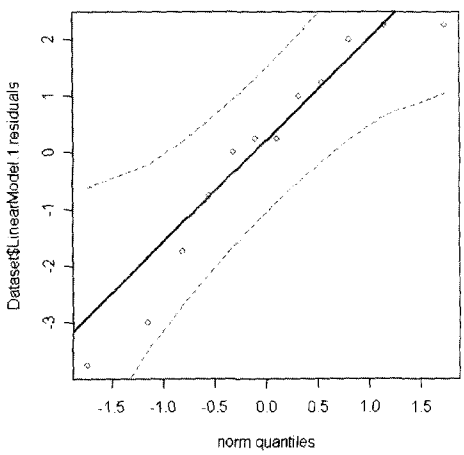


図15　残差の正規QQプロット

	A	B	fit	lwr	upr
1	A1	B1	20.75	17.65	23.85
2	A2	B1	23.75	20.65	26.85
3	A1	B2	23.75	20.65	26.85
4	A2	B2	26.75	23.65	29.85
5	A1	B3	13.00	9.90	16.10
6	A2	B3	16.00	12.90	19.10

図16　母平均の区間推定：fitは母平均、lwrは信頼下限、uprは信頼上限

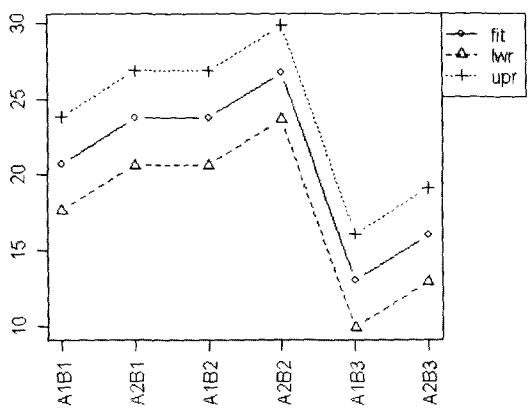


図17　母平均の区間推定のグラフ

6．おわりに

日本の品質管理においては、QCサークルから研究室レベルまで、様々な人たちが何らかのSQC・統計手法を実践している。そこでのRの利用を考えるとき、Rcmdrを用いたGUIでの利用がRの入門に適している。そのためにはRcmdrにおいて、日本的SQCのスタイルへの適合性を高めることが必要となる。現在、４節および５節で示したように、QC七つ道具から２元配置分散分析といった少し高度な手法までも、Rcmdrを通じてある程度簡単に適用することができる。区間推定や予測、それらのグラフ化を行うには少し、コマンドを入力する必要があるが、これらをRcmdrに実装するのはそれほど面倒ではない。さらに、３節で紹介した高度な手法も利用することができる。現在、Rcmdrを通じたRの品質管理での利用の整備を分散分析・実験計画法および多変量解析へと広げる作業を行っている。これらが完成すると、GUIでのほぼ完全なSQCパッケージを提供することができる。

参考文献

[１] 赤間世紀・山口喜博（2006）『Rによる統計入門』技報堂出版.

[２] 荒木孝治編著（2005）『フリーソフトウェアRによる統計的品質管理入門』日科技連出版.

[３] 岡田昌史編著（2004）『The R Book ―データ解析環境Rの活用事例集』九天社.

[４] Peter Dalgaard（2006）『Rによる医療統計学』丸善.

[５] 竹内俊彦（2005）『はじめてのS-PLUS/R言語プログラミング―例題で学ぶS-PLUS/R言語の基本』オーム社.

[６] 垂水共之・飯塚誠也（2006）『R/S-PLUSによる統計解析入門』共立出版.

[７] 椿広計（2006）『ビジネスへの統計モデルアプローチ』朝倉書店.

[８] 東京大学生物測定学研究室編（2004）『実践生物統計学―分子から生態まで』朝倉書店.

[９] 中澤港（2003）『Rによる統計解析の基礎』ピアソン・エデュケーション.

[10] 中村知靖・松井仁・前田忠彦（2006）『心理統計法への招待―統計をやさしく学び身近にするために』サイエンス社.

[11] 舟尾暢男（2005）『The R Tips ―データ解析環境Rの基本技・グラフィックス活用集』九天社.

[12] 舟尾暢男・高浪洋平（2006）『データ解析環境「R」―定番フリーソフトの基本操作から「グラフィックス」「統計解析」まで』工学社.

[13] 牧厚志・和合肇・西山茂・人見光太郎・吉川肇子・吉田栄介・濱岡豊（2005）『経済・経営のための統計学』有斐閣.

[14] 間瀬茂・神保雅一・鎌倉稔成・金藤浩司（2004）『工学のためのデータサイエンス入門―フリーな統計環境Rを用いたデータ解析 』数理工学社.

[15] U. リゲス（2006）『Rの基礎とプログラミング技法』シュプリンガー・ジャパン.

[16] 渡辺利夫（2005）『フレッシュマンから大学院生までのデータ解析･R言語』ナカニシヤ出版.

[17] Berger, R.L. (2007). Nonstandard operator precedence in Excel, *Computational Statistics & Data Analysis* 51, 2788-2791.

[18] Knusel, L. (2005). On the accuracy of statistical distributions in Microsoft Excel 2003, *Computational Statistics & Data Analysis* 48, 445-449.

[19] McCullough, B.D. and B. Wilson (2005). On the accuracy of statistical procedures in Microsoft Excel 2003, *Computational Statistics & Data Analysis* 49, 1244-1252.

[20] R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[21] Scrucca, L. (2004). qcc: An R package for quality control charting and statistical process control, *R News* 4, 1, June 11-17.