qistat パッケージ

数学 I「データの分析」用マクロ

北川 弘典

2015 年 1 月 24 日

このパッケージは，高等学校「数学 I」に出てくる各種統計量を TEX の内部だけで計算するパッケージで，`https://github.com/h-kitagawa/qistat` でひっそりと公開している．本来は，こういう統計処理は外部ソフトウェアを使うべきであり，本パッケージのように TEX だけで行うのは**余興**の域を出ないであろう．

1 動作環境

TEX 本体の数値計算機能をそのまま使うのはあまりにも貧弱である[*1]ので，数値計算には **fp パッケージ**[*2]を使用している．それ以外の外部パッケージは必要ないはずである．なお，plain TEX では使用できず，LATEX 2_ε の下で動作する．

2 命令一覧

本パッケージの提供する命令は基本的には `QI` で始まるが，統計量の出力命令には小文字の `qi` で始まるバリエーションがある．

2.1 初期化・値の追加

`\QIinit`⟨*name*⟩　⟨*name*⟩ という名称の内部リストを宣言する．初期値は空である．

`\QIappend`⟨*name*⟩⟨*data*⟩　内部リスト ⟨*name*⟩ の末尾に 1 つの値 ⟨*data*⟩ を追加する．標準では，データを追加した時点で次節に述べる各種統計量の再計算を行う．`\QIappend` 等を複数回連続で呼び出すときは無駄になってしまうので，その場合は `\QIappend*` と末尾にアスタリスクをつけることで再計算を抑止できる．

`\QIconcatI`⟨*name*⟩⟨*name2*⟩　内部リスト ⟨*name*⟩ の末尾に，内部リスト ⟨*name2*⟩ の内容を連結する．内部リスト ⟨*name*⟩ は予め宣言されていなくてはならない．

[*1] 2^{-16} pt を単位とした符号付き 31 bit の整数演算，言い換えれば整数部 14 bit，小数部 16 bit（及び符号ビット）の固定小数点演算．また基本的には四則演算の機能しかなく．Newton 法などで平方根計算をしてもあまり精度が出ない．

[*2] 整数部・小数部ともに 10 進 18 桁の範囲の固定小数点演算を実現させる Michael Mehlich 氏作のパッケージである．万が一導入されていなければ，`http://www.ctan.org/tex-archive/macros/latex/contrib/fp/` からダウンロードできる．

\QIconcatI* を末尾にアスタリスクをつけることで再計算を抑止できる．

\QIconcatC⟨*name*⟩{⟨*comma-separated list*⟩}　内部リスト ⟨*name*⟩ の末尾に，第 2 引数で与えられたコンマ区切りリストの内容を連結する．内部リスト ⟨*name*⟩ は予め宣言されていなくてはならない．\QIconcatC* を末尾にアスタリスクをつけることで再計算を抑止できる．

\QIdata⟨*name*⟩{⟨*comma-separated list*⟩}　内部リスト ⟨*name*⟩ の内容を第 2 引数で与えられたコンマ区切りリストで初期化する．内部リスト ⟨*name*⟩ が未宣言のときは，ここで宣言される．

\QIcopy⟨*name*⟩⟨*newname*⟩　内部リスト ⟨*name*⟩ の内容を内部リスト ⟨*newname*⟩ にコピーする．コピー先の ⟨*newname*⟩ が予め宣言されている必要はない．

\QIrecalc⟨*name*⟩　内部リスト ⟨*name*⟩ についての次節で述べる統計量を再計算する．この命令は，\QIconcatC*, \QIconcatI*, \QIappend* を実行した後以外は明示的に実行する必要はない．

コンマ区切りリスト中に「空の項目」はあってはならない．従って，

```
\QIdata{hoge}{1,2,3,4,}
```

のように最後にコンマを余計につけることも禁止である．

2.2 統計量の出力

\QIcnt⟨*name*⟩　内部リスト ⟨*name*⟩ のデータ数

\QImodemult⟨*name*⟩　内部リスト ⟨*name*⟩ において，最頻値が何回出現しているか

以下の命令の結果は整数とは限らないので，それぞれに対して，\QIavg に対して \qiavg のように最初の 2 文字が小文字になった変種が存在する．両者の違いは出力フォーマットで，

- 大文字の \QIavg 他は小数第 \value{QIdigit}位まで[*3]出力する．カウンタ QIdigit のデフォルト値は 2.
- 小文字の \qiavg 他は fp パッケージによる計算結果そのままで出力する．

例えば，

```
\QIdata{hoge}{1,0,0,0,0,0,0}\QIavg{hoge}, \qiavg{hoge}
```

からは，「0.14, 0.142857142857142857」が得られる．

\QIsum⟨*name*⟩　内部リスト ⟨*name*⟩ の合計値

\QIavg⟨*name*⟩　内部リスト ⟨*name*⟩ の平均値

\QIvar⟨*name*⟩　内部リスト ⟨*name*⟩ の分散（分母はデータ数 n）

\QIstdev⟨*name*⟩　内部リスト ⟨*name*⟩ の標準偏差

[*3] 元の値が整数などの場合は，小数点以下を補って出力する．

`\QImedian`⟨*name*⟩　内部リスト ⟨*name*⟩ の中央値

`\QIlqt`⟨*name*⟩　内部リスト ⟨*name*⟩ の第 1 四分位数 (lower quartile)

`\QIuqt`⟨*name*⟩　内部リスト ⟨*name*⟩ の第 3 四分位数 (upper quartile)

`\QIiqr`⟨*name*⟩　内部リスト ⟨*name*⟩ の四分位範囲 (IQR)

`\QImin`⟨*name*⟩　内部リスト ⟨*name*⟩ の最小値

`\QImax`⟨*name*⟩　内部リスト ⟨*name*⟩ の最大値

`\QImode`⟨*name*⟩　内部リスト ⟨*name*⟩ の最頻値（のうち最小のもの）

`\QImodemult`⟨*name*⟩　内部リスト ⟨*name*⟩ の最頻値が何回出現しているか

`\QIpct`⟨*name*⟩⟨*q*⟩　内部リスト ⟨*name*⟩ の q パーセンタイル[*4](percentile)

`\QIcov`⟨*name1*⟩⟨*name2*⟩　2 つの内部リスト ⟨*name1*⟩, ⟨*name2*⟩ の共分散 (covariance)

`\QIpcc`⟨*name1*⟩⟨*name2*⟩　2 つの内部リスト ⟨*name1*⟩, ⟨*name2*⟩ の相関係数
(Pearson product-moment correlation coefficient)

- `\QIlqt`, `\QIuqt` で計算する四分位数の定義は（少なくとも数研出版の）教科書で採用されているものを採用している．具体的には，第 1 四分位数は「下位のデータの中央値」，第 3 四分位数は「上位のデータの中央値」として計算する[*5]．
  Excel の quartile 関数による値と同じ計算法を用いたいならば，`\QIpct`⟨*name*⟩`{25}`, `\QIpct`⟨*name*⟩`{75}`などとすること．
- 内部リスト ⟨*name*⟩ が空リストの場合でも，上記の命令類はエラーは返さず，適当な値となる．例えば `\QImin`, `\QImax` はそれぞれ `fp` パッケージの扱える最大値・最小値となる．
- 本節の命令で出力される統計量のうち内部リストの変更に伴い自動的に計算されるだけであり，**命令実行時にその都度計算されるわけではない**．従って，前節で述べた `\QIconcatC*`, `\QIconcatI*`, `\QIappend*` の 3 命令を用いてリストにデータを一通り追加した場合は，`\QIrecalc` を手動で実行する必要がある．

2.3 その他

`\QIout`⟨*name*⟩, `\qiout`⟨*name*⟩　内部リスト ⟨*name*⟩ の内容を「`,`␣」で区切って出力．
前節と同様に，小文字の `\qiout` は各データをそのまま出力するが，大文字の `\QIout` は各データを丸めて出力する．

`\QIshift`⟨*name*⟩, `\qishift`⟨*name*⟩　内部リスト ⟨*name*⟩ の先頭のデータを出力する．⟨*name*⟩ からは先頭のデータは取り除かれる．
前節と同様に，小文字の `\qishift` は各データをそのまま出力するが，大文字の `\QIshift` は各データを丸めて出力する
またこの命令は，内部リストの内容を表組などの整形用途を目的として作成されて

[*4] データ数を n としたとき，$1+q(n-1)/100$ 番目のデータ値，として計算している．整数でなかった場合は隣り合ったデータから線型に補完．

[*5] データ数が $2k+1$ の場合，「下位のデータ」「上位のデータ」は中央値を含めない k 個のデータ達のことを指す．Tukey のヒンジ値というそうだ．

おり，「先頭のデータを取り除く」操作はグローバルなものであり，さらに統計量の再計算は行わない．

`\QIsort`⟨*name*⟩⟨*newname*⟩　内部リスト ⟨*name*⟩ 内のデータを小さい順に並べた内部リストを ⟨*newname*⟩ として参照できるようにする．⟨*newname*⟩ が予め宣言されている必要はなく，また，⟨*name*⟩ が更新されても ⟨*newname*⟩ の値は追従しない．

`\QIboxplot[`⟨*height*⟩`]`⟨*list*⟩　⟨*list*⟩ 内のデータを元に箱ひげ図を作成する．`picture` 環境の内部で

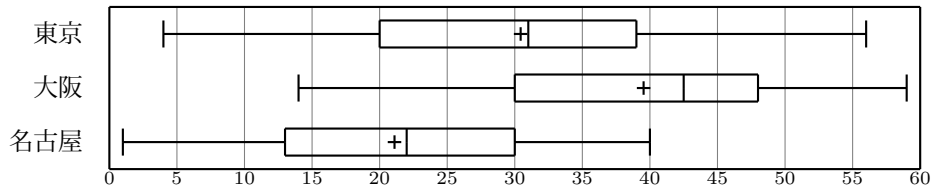
```
\put(0,2){\QIboxplot{tokyo}}
```

のように用いることを想定しており，`\put` の座標を 0 として，`\unitlength` を単位長さとし，右方向を正として組まれる．箱は高さが 2⟨*height*⟩`\unitlength`，最大値・最小値の位置の縦棒と平均値の位置の十字の大きさはその半分の ⟨*height*⟩`\unitlength` としている．省略時の ⟨*height*⟩ は 1 である．

3　例 1

元データは 1984–2013 の熱帯夜[*6]日数．気象庁ホームページの `http://www.data.jma.go.jp/obd/stats/etrn/index.php` より．

X	$\overline{X}$	σ^2	σ	min	Q_1	中央値	Q_3	max
東京	30.43	144.65	12.03	4	20	31	39	56
大阪	39.53	132.98	11.53	14	30	42.5	48	59
名古屋	21.10	99.76	9.99	1	13	22	30	40



	東京・大阪	東京・名古屋	大阪・名古屋
共分散	107.37	90.96	100.95
相関係数	0.77	0.76	0.88

4　例 2

4, 3, 3, 1, 3, 2, 1, 6, 6, 5, 2, 4, 4, 1, 5, 1, 6, 5, 5, 1, 1, 3, 6, 2, 4, 2, 2, 1, 6, 2, 4, 3, 6, 1, 1, 4, 5, 4, 5, 4, 6, 6, 1, 3, 5, 2, 1, 5, 2, 4, 3, 5, 1, 2, 2, 5, 6, 2, 2, 6, 3, 4, 1, 6, 3, 3, 2, 3, 6, 6, 6, 1, 2, 4, 2, 3, 3, 6, 5, 3, 2, 5, 5, 4, 2, 2, 1, 1, 2, 1, 5, 3, 1, 1, 1, 2, 1, 5, 1, 2, 6, 4, 4, 2, 5, 1, 4, 2, 6, 6, 4, 6, 5, 4, 1, 6, 6, 1, 1, 2, 1, 4, 3, 6, 3, 2, 6, 3, 6, 2, 3, 3, 3, 3, 4, 4,

[*6] 熱帯夜とは，最低気温が 25 度以上の日のことである．

5, 5, 2, 4, 2, 4, 5, 6, 4, 5, 5, 5, 5, 4, 4, 6, 3, 1, 4, 2, 5, 2, 4, 2, 1, 2, 1, 4, 6, 4, 2, 2, 4, 4,
3, 3, 1, 6, 3, 2, 3, 2, 3, 5, 6, 3, 1, 4, 4, 6, 4, 2, 2, 3, 5, 3, 2, 5, 3, 6, 1, 1, 5, 1, 5, 1, 1, 3,
4, 4, 4, 6, 4, 4, 4, 2, 6, 5, 1, 2, 6, 6, 5, 4, 4, 6, 5, 2, 5, 4, 1, 5, 5, 5, 5, 2, 4, 5, 5, 3, 3, 3,
6, 4, 6, 4, 1, 4, 3, 2, 4, 6, 1, 2, 5, 6, 3, 3, 6, 3, 2, 4, 4, 3, 6, 6, 2, 5, 6, 5, 2, 1, 5, 5, 4, 1

という 272 個のデータにおける最頻値（のうち最小のもの）は 4 で，51 回出現している．

5 \QIlqt・\QIuqt と \QIpct

5.1 データ数が $4k+1$ のとき

元データ（test）が 1, 2, 3, 4, 5, 6, 7, 8, 9 のとき，

\QIlqt{test} = 2.50, \QIuqt{test} = 7.50
\QIpct{test}{25} = 3.00, \QIpct{test}{75} = 7.00.

5.2 データ数が $4k+2$ のとき

元データ（test）が 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 のとき，

\QIlqt{test} = 3.00, \QIuqt{test} = 8.00
\QIpct{test}{25} = 3.25, \QIpct{test}{75} = 7.75.

5.3 データ数が $4k+3$ のとき

元データ（test）が 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 のとき，

\QIlqt{test} = 3.00, \QIuqt{test} = 9.00
\QIpct{test}{25} = 3.50, \QIpct{test}{75} = 8.50.

5.4 データ数が $4k$ のとき

元データ（test）が 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 のとき，

\QIlqt{test} = 3.50, \QIuqt{test} = 9.50
\QIpct{test}{25} = 3.75, \QIpct{test}{75} = 9.25.