Graphical Data and Data Graphics

Paul Murrell

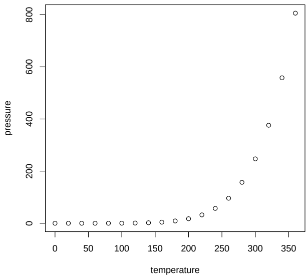
The University of Auckland

July 12 2007

Graphical Statistics

```
> pressure
  temperature pressure
1           0   0.0002
2          20   0.0012
3          40   0.0060
4          60   0.0300
5          80   0.0900
6         100   0.2700
7         120   0.7500
8         140   1.8500
...
```

⟶



Statistical Graphics

```
> pressure
  temperature pressure
1           0   0.0002
2          20   0.0012
3          40   0.0060
4          60   0.0300
5          80   0.0900
6         100   0.2700
7         120   0.7500
8         140   1.8500
...
```

→

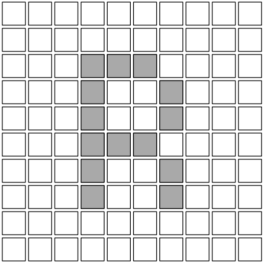
Graphical Data and Data Graphics

- Graphical Statistics: *data* → *plot*
- Statistical Graphics: *data* → *plot*

- Graphical Data: *plot* → *data*
- Data Graphics: *plot* → *data*

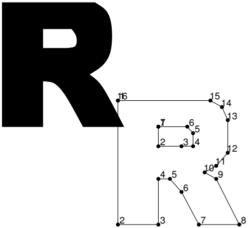
Graphical Formats

Raster



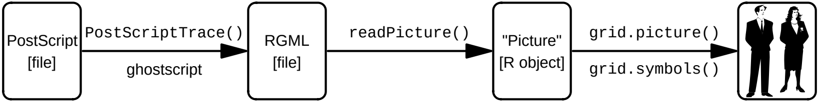
pixmap package
EBimage package

Vector



grImport package

The grImport Package

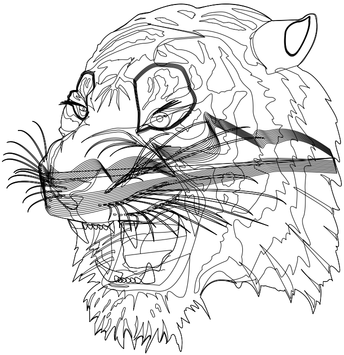


The PostScript Bezier Tiger

```
%!PS-Adobe-2.0 EPSF-1.2
%%Creator: Adobe Illustrator(TM)
%%For: OpenWindows Version 2
%%Title: tiger.eps
...
.8 setgray
clippath fill
-110 -300 translate
1.1 dup scale

0 g
0 G
0 i
0 J
0 j
0.172 w
10 M
[]0 d
0 0 0 0 k
...
```

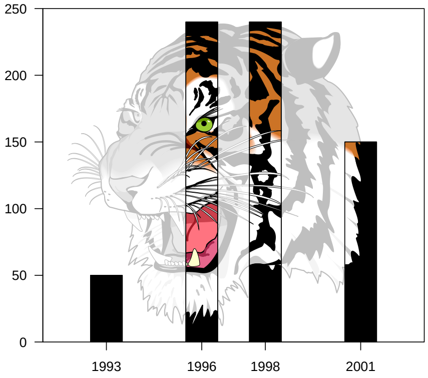
Converting the Tiger to Data

```
PostScriptTrace("tiger.ps")

tiger <-
  readPicture("tiger.ps.xml")
```



Using the Tiger in a Plot

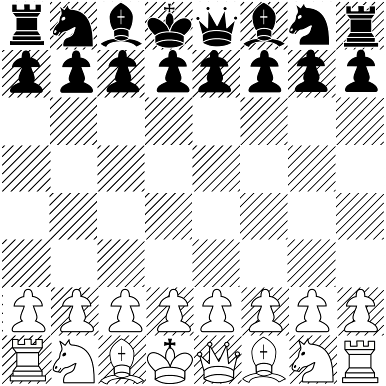
```
grid.picture(tiger)
```



A Chess Board

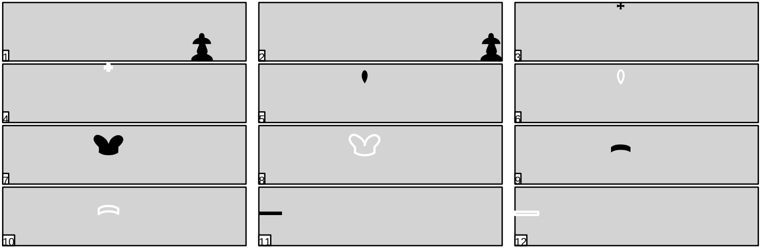
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG"
"http://www.w3.org/TR/2001/REC-SVG...">
<!-- Created with Sodipodi -->
<svg version="1.0">
...
  <g
     style="font-size:12;"
     id="g874">
    <path
       d="M 0 437 L 437 0 "
       style="fill:none;fill-opacity:1"
       id="path616" />
...
```

```
# Convert SVG to PostScript
# using InkScape

PostScriptTrace("chess.ps")

chess <-
  readPicture("chess.ps.xml")
```



The Paths in the Chess Board

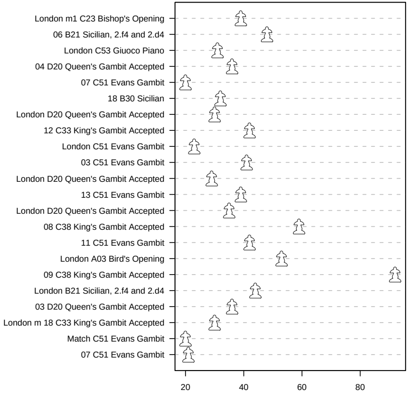
```
picturePaths(chess[125:136])
```



A Chess Piece as a Plotting Symbols

The number of moves required to complete chess games for different opening gambits. From the career of Louis Charles Mahe De La Bourdonnais (circa 1830).

```
grid.symbols(
  chess[205:206],
  x=games$num.moves,
  y=1:ngames,
  "native",
  size=unit(0.5, "cm"))
```



Statistical Data Graphics

- Graphical Statistics: *data* → *plot*
- Statistical Graphics: *data* → *plot*

- Graphical Data: *plot* → *data*
- Data Graphics: *plot* → *data*

- Statistical Data Graphics: *data* → *plot* → *data*

A Published Plot

erpho
Eastern Region Public Health Observatory
NHS
November 2004 Issue 4 ISSN 1477-7290

INphoRM

Presenting performance indicators: alternative approaches

Information on Public Health Observatory recommended methods

Introduction

Measurement of performance in the NHS involves the collection, analysis and presentation of data in the form of performance indicators. While data *analysis* is usually carried out by individuals with specific technical skills, data *collection* is often the responsibility of clinicians and managers. Moreover, *interpretation* of the resulting indicators is open to anyone including patients, journalists, politicians, civil servants and managers. Many of these people do not always have a detailed understanding of the technical issues underlying the collection and presentation of indicator data. It is therefore important that indicators are both accurate and presented in a way that does not result in unfair criticism or unjustified praise.

This issue of *INphoRM* provides technical information about improved approaches to presenting indicators. The first part looks at *process control charts* and *funnel plots* and the second part introduces *cumulative failure* and *cumulative summation graphs*. The techniques described are supported with example spreadsheets available from the erpho website (see 'Further resources'). More general information about the principles of measuring performance can be found in *INpho* issue 4, 'Quantifying performance: using performance indicators'[1].

Current methods

Analysis
Most indicators are constructed, interpreted, and analysed using a standard approach. The measurement itself is made up of a numerator and a denominator. The resulting proportion or rate can then be compared with a standard (e.g. a regional average or a predetermined benchmark).

Measurement:	Comparator:
a (numerator) / **b** (denominator)	**c** (e.g. average or benchmark)

Statistical tests may be used to determine *how significant is the difference* between the measurement and the comparator.

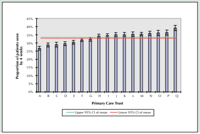
Presentation
The results of such analyses are usually presented as ranks or league tables, often using 'traffic light' coding (green for satisfactory performance, amber when there is some concern and red for unsatisfactory performance). A Primary Care Trust (PCT) star rating in the NHS, itself an important indicator with important consequences, is a composite of other individual performance measures.

Figure 1 shows PCTs in Norfolk, Suffolk and Cambridgeshire ranked according to the proportion of their patients referred to hospital that are seen within four weeks. However, ranking in this way has severe limitations and great potential for misinterpretation.

Limitations of current methods

Methods based on ranking, such as league tables or percentiles, have a number of flaws. The main problem with ranking is the implicit assumption that there is *any* performance difference between organisations. Simply because institutions may produce different values for an indicator, and we naturally tend to rank these values, does not mean that we are observing variation in performance. *All* systems within which institutions operate, no matter how stable, will produce variable outcomes. The questions we need to answer are: Is the observed variation more or less than we would normally expect? Are "poor performers" genuine outliers? Are there exceptionally good performers? And so on.

Figure 1.
Four-week waiting by PCT: PCTs (identified by letter) are ranked according to the proportion of their referrals seen by 4 weeks.
Source: QM08 returns for Quarter 3, 2002-2003, Department of Health (no longer online).



Ranking fails to allow for the variation associated with measurement that occurs even in the most stable system.

Ranking fails to allow for the variation associated with measurement that occurs even in the most stable system.[2] This failure to allow for insignificant and meaningless variation leads to ranking being invalid. A good example of this was the ranking of the 15 English Hospital Trusts with the lowest mortality rates by Dr Foster (an independent organisation that produces a Good Hospital Guide).[3] In order to show the uncertainty of the rankings, Dr Foster also presented the probability for each Trust that its place in the rankings was correct. Two out of fifteen Trusts had a probability of less than 60% of being in the top ranks. Using confidence intervals to indicate the range of uncertainty can help the reader towards a better interpretation, but it doesn't solve the problem:

- There is a natural tendency to focus on the position of an organisation in a table and ignore the confidence interval.
- The comparison of multiple confidence intervals is a form of multiple significance testing that can lead to serious misinterpretation. (Remember that on average 1 in every 20 measurements will fall outside the 95% confidence intervals purely by chance.)
- Confidence intervals are not readily understood by everyone who uses performance data.

A critique of the weaknesses of rank-based approaches can be found in a recent paper on public sector performance indicators from the Royal Statistical Society.[4]

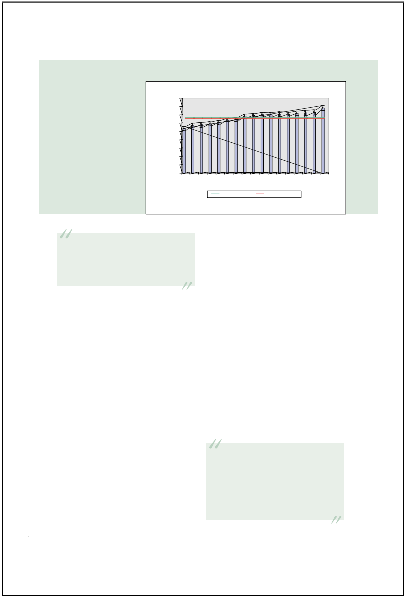
An alternative approach

Rather than assuming a performance difference between organisations, a different approach is to begin by assuming that they are all part of a single health care system, and examining the degree of variation observed with that expected.[5] Well-tried techniques such as 'statistical process control' can then be used to distinguish between those parts of the system that are operating within normal limits and those parts that show greater than expected variation. These techniques involve plotting data on a scatter plot and then superimposing 'control limits' onto the graph. The control limits divide those points *between* the control limits (which exhibit 'common-cause' variation) from those points lying *outside* the control limits (which exhibit 'special-cause' variation). Common-cause variation is the variation inherent within any system, and can never be completely eliminated. Special-cause variation *cannot* be attributed to the inherent variability within a system and requires further explanation to identify its cause. Once an explanation has been identified it should be possible to correct special-cause variation through appropriate changes.

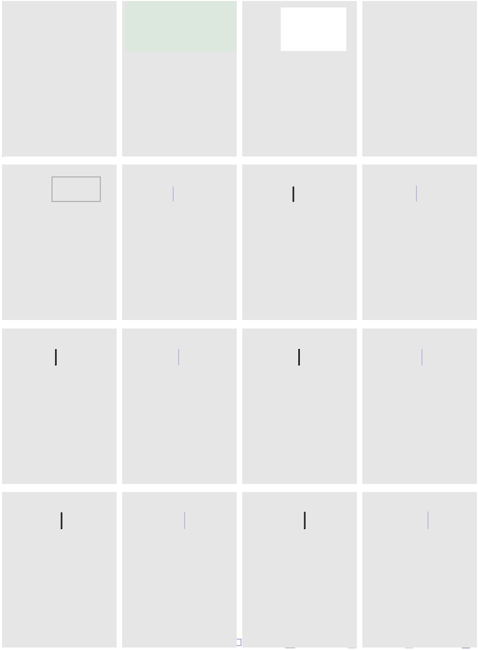
In effect, a process control chart allows organisations, on the basis of their performance data, to be split into three groups: those whose performance is unremarkable and as expected (the majority of organisations in a stable system),

The control limits divide those points between the control limits (which exhibit 'common-cause' variation) from those points lying outside the control limits (which exhibit 'special-cause' variation).

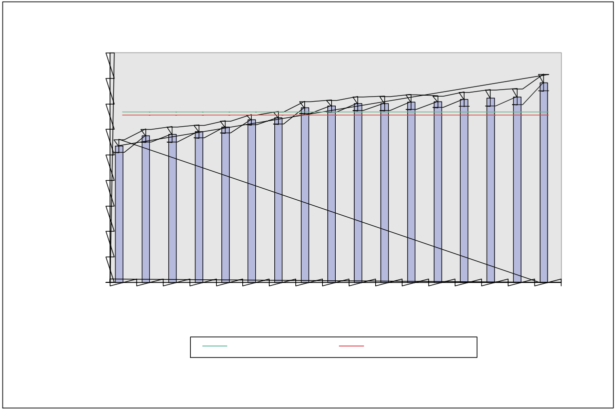
```
# Extract just page 2
# and convert to PostScript

PostScriptTrace("Fig1.ps")

Fig1 <-
  readPicture("Fig1.ps.xml")

grid.picture(Fig1)
```



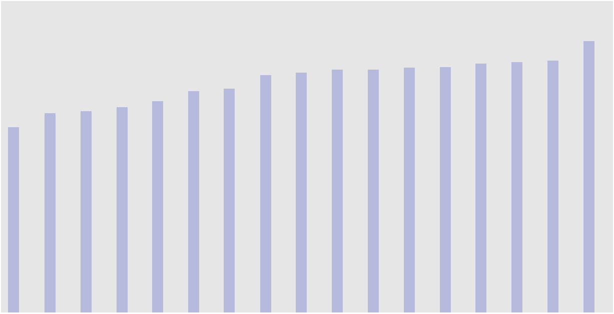
`picturePaths(Fig1)`



```
grid.picture(Fig1[4:48])
```



```
> barePlot <- Fig1[seq(4, 38, 2)]

> grid.picture(barePlot)
```

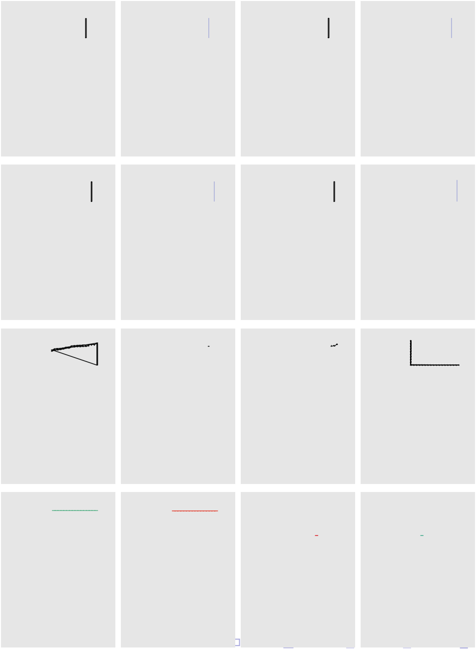


```
> slotNames(barePlot)
[1] "paths"   "summary"
> barePlot@summary
An object of class "PictureSummary"
Slot "numPaths":
[1] 18

Slot "xscale":
[1] 2563 5046

Slot "yscale":
[1] 6108 7371
```

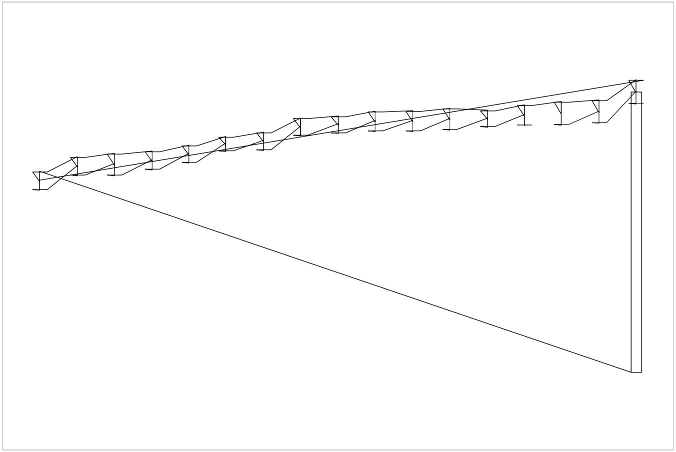
```
> class(barePlot@paths)
[1] "list"
> barePlot@paths[[1]]
An object of class "PictureFill"
Slot "x":
move line line line line
2563 5046 5046 2563 2563

Slot "y":
move line line line line
6109 6109 7371 7371 6109

Slot "rgb":
[1] "#E6E6E6"

Slot "lwd":
[1] 1.33
```

```
> scaledMax <- function(x, summary) {
      (max(x@y) - summary@yscale[1]) /
      diff(range(summary@yscale))
  }
> barProportions <- sapply(barePlot@paths[-1],
                           scaledMax,
                           barePlot@summary)
> barProportions * 45
 [1] 26.8 28.8 29.1 29.6 30.5 31.9 32.3 34.3 34.6 35.1 35.1
[12] 35.4 35.5 35.9 36.2 36.4 39.2
```

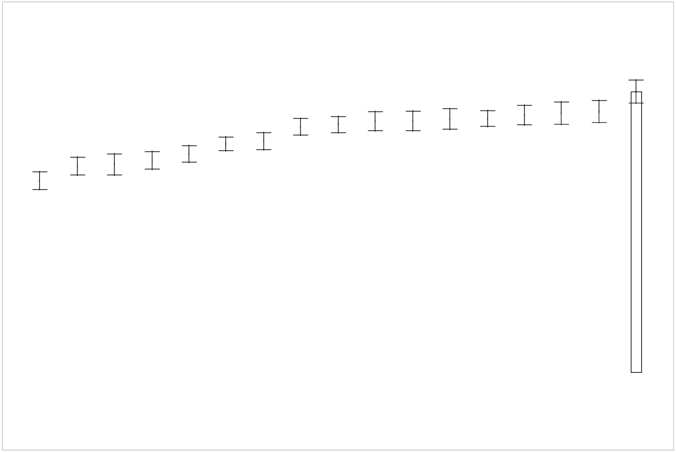
`picturePaths(Fig1)`



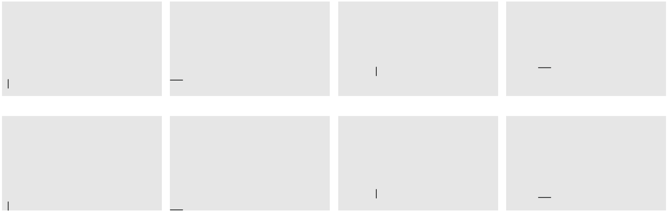
```
> grid.picture(Fig1[39:41])
```



```
> errorBars <- explodePaths(Fig1[39:41])
> grid.picture(errorBars)
```



```
> picturePaths(errorBars)
```



```
> topBars <- errorBars[seq(3, 35, 2)]
> bottomBars <- errorBars[seq(37, 69, 2)]
> scaledMin <- function(x, summary) {
      (min(x@y) - summary@yscale[1]) /
       diff(range(summary@yscale))
  }
> barMaxProp <- sapply(topBars@paths,
                       scaledMax,
                       barePlot@summary)
> barMinProp <- sapply(bottomBars@paths,
                       scaledMin,
                       barePlot@summary)
```

```
> barMaxProp * 45
 [1] 28.0 30.0 30.5 30.8 31.6 32.8 33.4 35.4 35.7 36.3 36.4
[12] 36.8 36.5 37.2 37.7 37.9 40.8
> barMinProp * 45
 [1] 25.5 27.5 27.5 28.4 29.3 30.9 31.1 33.1 33.4 33.7 33.7
[12] 33.9 34.3 34.5 34.6 34.8 37.6
```

Graphical Data Graphical Statistics

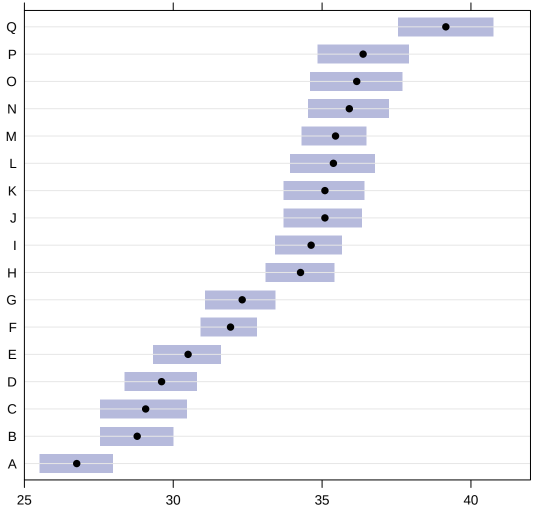
- Graphical Statistics: *data* $\longrightarrow$ *plot*
- Statistical Graphics: *data* $\longrightarrow$ *plot*

- Graphical Data: *plot* $\longrightarrow$ *data*
- Data Graphics: *plot* $\longrightarrow$ *data*

- Statistical Data Graphics: *data* $\longrightarrow$ *plot* $\longrightarrow$ *data*

- Graphical Data Graphical Statistics:
  data $\longrightarrow$ *plot* $\longrightarrow$ *data* $\longrightarrow$ *plot*

```
dotplot(LETTERS[1:17] ~ barProportions*45)
```



Acknowledgements

- The tiger image is part of the **ghostscript** distribution; the tiger data are from `http://www.globaltiger.org/population.htm`.
- The greyscale version of the tiger used the **colorspace** package by **Ross Ihaka**.
- The chess board image (by **Jose Hevia**) is from the **Open Clip Art Library** `http://openclipart.org/clipart//recreation/games/chess/chess_game_01.svg`
- The chess data are from **chessgames.com** `http://www.chessgames.com/perl/chess.pl?page=1&pid=31596`
- **INphoRM** (Information on Public Health Observatory recommended methods) is a publication of the **Eastern Region Public Health Observatory**.
- The idea of extracting the data from a plot in an issue of INphoRM came from **Ted Harding**.