Box Plots and Scatter Plots with TEX Macros

A.J. Van Haagen

1 Introduction

The powerful macro facilities of the TEX language can be employed to create simple sets of instructions for drawing box plots and scatter plots. Both are important tools in the graphical analysis of statistical data.

Box plots, more fully known as box-and-whisker plots, were introduced by Tukey [2] as a graphical tool to give a summary of the distribution of a set of data by five numbers: the lower extreme, the first quartile, the median, the third quartile, and the upper extreme. An example of a box graph is given in Figure 1. The code for Figure 1 shows how easy it is to draw such plots with the set of definitions in the file `macboxplot.tex`. The figure shows how box plots can be used to compare batches of data. For a more detailed discussion of box plots see [3] and the references cited there.

Scatter plots serve many useful purposes in statistics. They are pervasive in regression analysis and can also be used effectively in areas like nonparametric statistics and Time Series Analysis. With the definitions in the file `macplot.tex` scatter plots and also some related graphs can be drawn very easily. An illustration is provided by Figure 2. The input has been computed by SAS [1].

2 Drawing Box Plots with TEX Macros

To draw box plots, first input the macros of the file `macboxplot.tex` with the `\input macboxplot.tex` command. This command is followed by the optional scale commands `\xscale[`*scalefactor*`]` and `\yscale[`*scalefactor*`]`. The entry *scalefactor* is an integer in the range $1, \ldots, 100$. The default value of the `\xscale` and `\yscale` command is 100. These commands can be used to scale the picture in horizontal and vertical direction, respectively.

The commands for drawing the box plots open with `\beginboxplot` and close with `\endboxplot`. The `\beginboxplot` command must be followed by the command `\range[`*low*`,`*high*`]`. The entries *low* and *high* should be integers greater than -2^{31}, but less than 2^{31}, such that $high - low \leq 214748$. The entry *low* corresponds with the bottom of the rectangular box in which the box plots are enclosed, the entry *high* with the top. The actual choice depends on the box plot parameters and will be discussed below.

Next comes the command `\ordinates[`*l*`,`*h*`] [`*n*`,`*t*`,`*lpos*`,`*hpos*`]` which provides ordinates and tick marks. Here l and h are integers greater than -2^{30}, but less than 2^{30}, such that $l < h$ and $h - l \leq 214748$, n is the number of tick marks ($n \geq 2$), and t is 1,10,100,1000, or 10000. The parameters *lpos* and *hpos* are optional. They determine the positions of the first and last tick mark with their corresponding ordinate values, respectively. The values *lpos* and *hpos* must be integers between *low* and *high*. Their default values are *low* and *high*, respectively. For correct results $h - l$ should be a multiple of $n - 1$. The values l/t, $(l + i)/t$, $(l + 2i)/t$, $\ldots$, h/t, where $i = (h - l)/(n - 1)$, are placed to the left of the tick marks. Small tick marks are drawn half way between the ones just referred to.

A box plot is drawn by the command:

`\boxplot[`*boxplotlow*`,`*1st quartile*`,`*median*`,`
3rd quartile`,`*boxplothigh*`,`*mean*`]`,

where *boxplotlow* is the lower extreme of the box plot and *boxplothigh* is the upper extreme. The entry *mean* is optional. The mean is indicated in the box plots by an asterisk. All entries are integers between *low* and *high*. To obtain these entries multiply the box plot parameters, as computed by SAS for instance, by a suitable power of 10 and round off or truncate. Convenient integer values for *low* and *high* are then selected.

The command `\boxplot` is followed by the optional commands `\outsiders[...]`, `\outlabelleft[...][...]`, and `\outlabelright[...][...]`. The number of box plots that can be drawn is limited only by the capacity of the TEX memory. They will be scaled and positioned automatically. The `\outsiders` command can handle any number of outside values, limited again only by the memory capacity of TEX. The command `\outlabelleft` writes a label to the left of an outside value, and `\outlabelright` writes one to the right.

After the box plots have been drawn there are the following options. The command `\boxplotlabels[...]` writes labels under the box plots. The labels may be numerical or may consist of text. The `\vertlabel[...]` command writes a vertical label to the left of the ordinates. Finally, the command `\text[...]` inserts text below the box plot labels.

The way these commands are used is evident from their appearance in Figure 1.

Since it is possible to scale down the picture to any desired size, one may wish the box graph to appear in the middle of text. The `\beginboxplot ... \endboxplot` sequence

should then be enclosed by the TEX commands `\midinsert ... \endinsert`.

The macros in the `macboxplot.tex` file are compatible with the LATEX system developed by Leslie Lamport.

3 Drawing Scatter Plots with TEX Macros

To draw a scatter plot, first input the macros of the file `macplot.tex` with the `\input macplot.tex` command. This command is followed by the optional scale commands `\xscale[...]` and `\yscale[...]` which are identical to the commands of the same name in the file `macboxplot.tex`.

The commands for drawing a scatter plot open with `\beginplot` and close with `\endplot`. The `\beginplot` command must be followed by the commands `\xrange[...]` and `\yrange[...]`. These commands set the lower and upper limits of the x coordinates and the y coordinates of the points to be plotted. The `\yrange` command is identical to `\range`. Next come the commands `\xaxis[...][...]` and `\yaxis[...][...]` which provide scales for the x axis and y axis. The command `\yaxis` is identical to `\ordinates`. Instead of these two commands, one can also apply the commands:

`\xlabels[(`x_1`,`*xlabel1*`)(`x_2`,`*xlabel2*`),...]`,
and
`\ylabels[(`y_1`,`*ylabel1*`)(`y_2`,`*ylabel2*`),...]`.

The first entry within each pair of parentheses indicates the location on the x axis and y axis, respectively. It should be an integer between the parameters in `\xrange` and `\yrange`, respectively. The second entry is the label one wants to write near that location.

The following commands: `\points[...][...]`, `\leftlabel[...][...]`, `\rightlabel[...][...]`, `\horlabel[...]`, `\vertlabel[...]`, `\hordash[...]`, and `\vertdash[...]` can be applied in any order.

The `\points[`*plottingsymbol*`][(`x_1`,`y_1`)(`x_2`,`y_2`),...,(`x_n`,`y_n`)]` command places the plotting symbol in the locations given by the coordinate pairs. The plotting symbol can be almost any symbol in the TEX system.

The `\leftlabel[...][...]` command writes a label to the left of a point and `\rightlabel[...][...]` writes one to the right. The `\horlabel[...]` command writes a label below the x axis and `\vertlabel[...]` writes one to the left of the y axis. The commands `\hordash[...]` and `\vertdash[...]` draw horizontal and vertical lines of dashes, respectively. The last command is the `\text[...]` command which writes text below the horizontal label.

The use of most of the these commands is illustrated in Figure 2. The macros are again compatible with the LATEX system.

Besides the TEX macros described above, there is a definition which requires PostScript for drawing slanted lines. It is of the form:

`\PSpolyline[`*linethickness*`][(`x_1`,`y_1`)(`x_2`,`y_2`),...(`x_n`,`y_n`)]`.

The entry *linethickness* is a number which sets the thickness of the polygonal line connecting the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The unit used by PostScript is the bp.

Acknowledgement. The author wishes to thank Professor L. Gordon for his suggestion to try to make TEX draw statistical graphs and for his constructive remarks. He also wishes to thank the department of mathematics of the University of Southern California for allowing him to use their excellent TEX facilities.

References

[1] SAS Institute Inc. *SAS/STAT Guide for Personal Computers, Version 6 Edition.* Cary, North Carolina, 1985.

[2] John W. Tukey. *Exploratory Data Analysis.* Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.

[3] Antonius J. Van Haagen. *Box Plots and Scatter Plots with TEX Macros.* A Thesis for the Degree of Master of Science in Statistics, University of Southern California, Los Angeles, 1987.

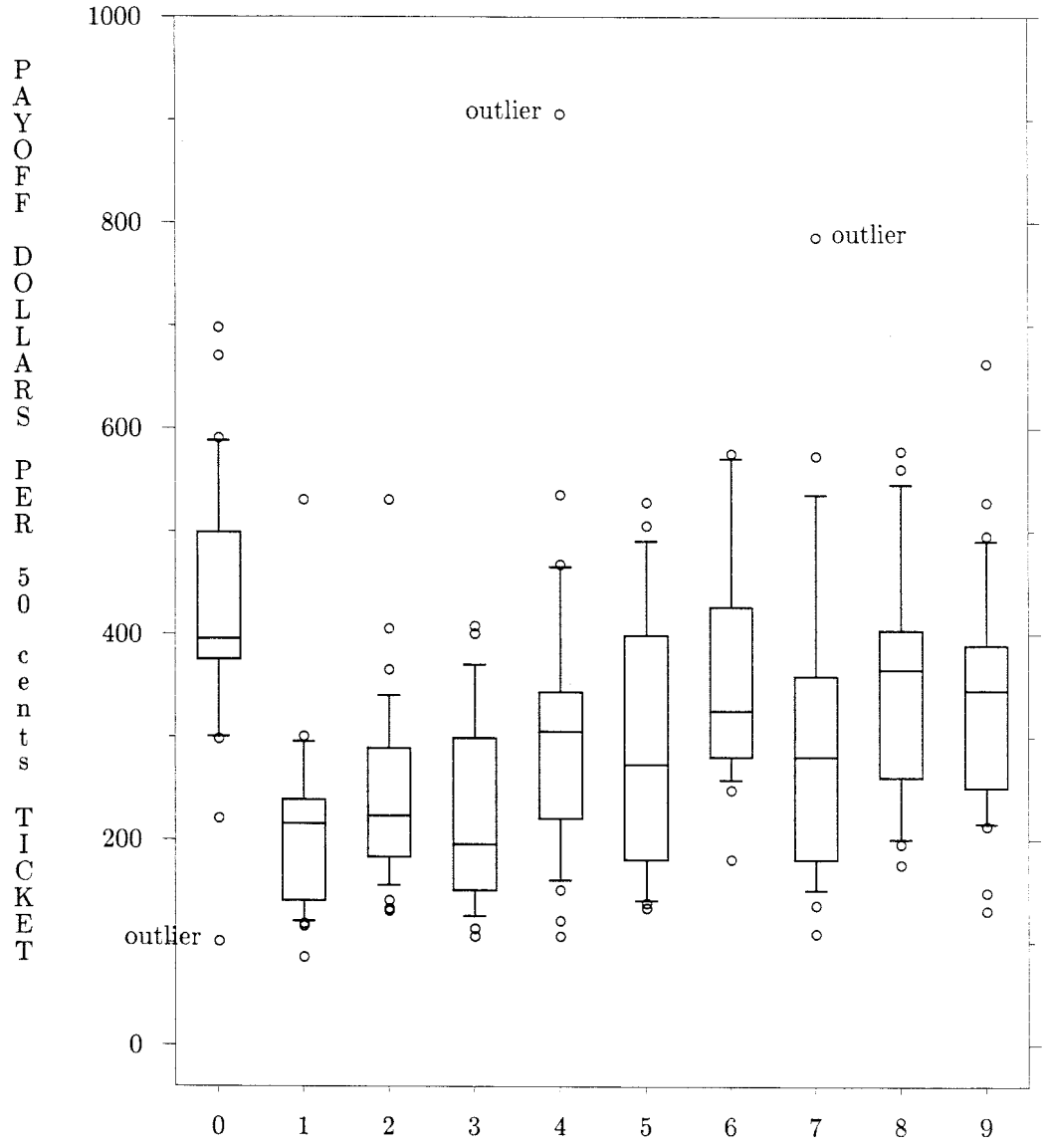

Figure 1: BOX GRAPH. The vertical scale is payoff of the New Jersey lottery, or numbers game, in which a player picks a three-digit number from 000 to 999. Winners share half of the pot. Each box graph shows the distribution of payoffs for all numbers with a particular leading digit. A leading digit of zero has the highest payoffs because fewer people tend to pick them. As the leading digit increases from one to nine the payoffs increase in a zigzag fashion, showing odd first digits are preferred to even. (From: The Elements of Graphing Data by William S.Cleveland. The data for this picture have been obtained by measuring the picture in Cleveland's book.)

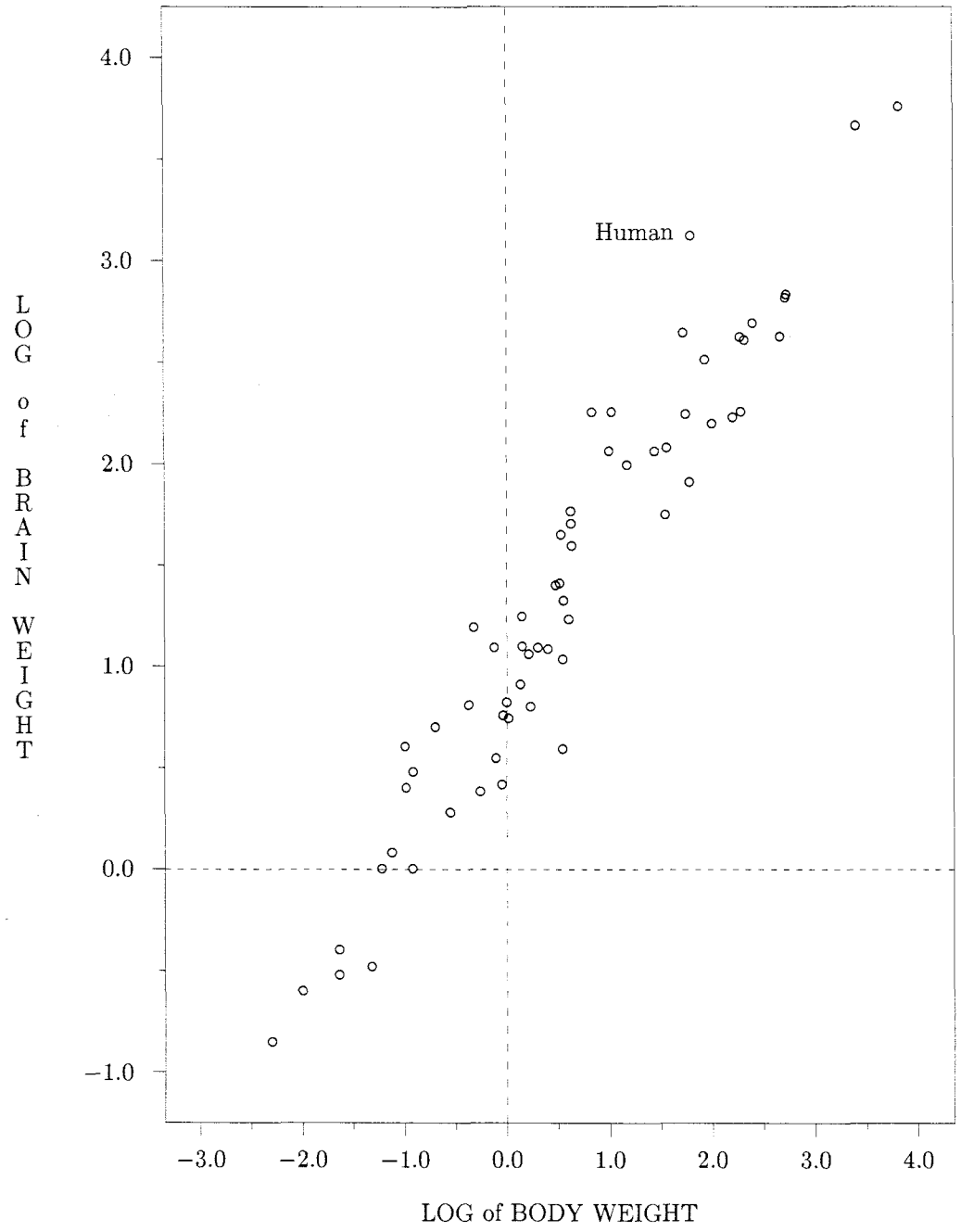

Figure 2: A scatter plot of brain weight data. The variable $log_{10}(brainweight)$ is plotted against the variable $log_{10}(bodyweight)$. Brain weight is the average brain weight in grams and body weight is the average body weight in kgs for 62 species of mammals.