Dynamic documents

David Allen

Abstract

The term *dynamic document* covers a wide assortment of documents; see Wikipedia for a general definition. The discussion here involves a situation where the document is a report including statistical analyses and graphical displays based on data. The data are continually being augmented or replaced. What is needed is a way to automate the revision of the document when the data changes. Our approach here is to use R to do the statistical calculations and graphics, tikzDevice (an R package) to output the graphics to LaTeX, and knitr (also an R package) to process an input file to a LaTeX file.

1 The Kentucky Senate race

On November 4, 2014, the Commonwealth of Kentucky will elect a United States Senator. The candidates are Alison Lundergan Grimes and Mitch McConnell. This race has high national impact and is closely watched. A poll yields the number of people in a sample, from a population of potential voters, favoring each candidate. The proportion of the sample favoring Alison (or Mitch) is reported. However, this provides no indication of the sampling variability.

The parameter of interest is the population proportion favoring Alison. A *credible interval* is such that the parameter lies within the interval with high probability. A credible interval is a more informative mode of presentation, as it conveys the uncertainty of knowledge about the parameter. Calculation of the credible interval is the statistical analysis portion of the report.

Denote the proportion of the population favoring Alison by p. The first step in calculating a credible interval is finding the posterior density function of p given the sample results. One needs to select a level of credibility. The value 0.95 has a strong tradition and is used here. The 0.95 credibility interval is an interval (p_1, p_2) where $P(p_1 < p < p_2) = 0.95$. There are multiple intervals satisfying the probability statement. The interval having minimal length is usually used.

An example assuming a sample with 55 favoring Alison and 45 favoring Mitch is shown in fig. 1. The 0.95 credible interval $(0.4528, 0.6428)$ is the base of the shaded region. A report might look like:

> A current poll produced 55 potential voters favoring Alison and 45 favoring Mitch. These results give a 0.95 credible interval for the proportion favoring Alison of $(0.4528, 0.6428)$.

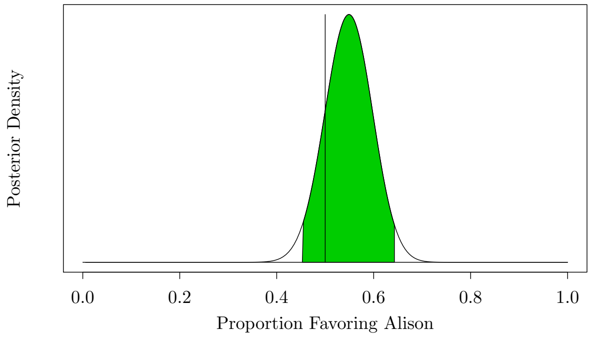

Figure 1: Graph showing credible interval.

The report might also include the above graph (fig. 1).

The preceding graphic and the credible interval were produced with R. The output from R was then transcribed to a LaTeX file to produce the "report". Polling will be a continuing activity from now until election day. Rerunning R and cutting and pasting output into a LaTeX document is tedious and error prone. Subsequent sections show how to automate the process.

2 Ti*k*Z graphics

Ti*k*Z is a graphics package used in conjunction with TeX. It is included with most distributions of TeX, or may be downloaded at `http://sourceforge.net/projects/pgf/`. A large selection of examples of Ti*k*Z graphics are posted at `http://www.texample.net/tikz/examples/`.

I think it likely that most *TUGboat* readers are familiar with *TikZ*. I give just two examples I have composed. The graphic in fig. 2 was hand-coded in Sketch (`http://www.frontiernet.net/~eugene.ressler/`), and then processed into Ti*k*Z, and fig. 3 was written directly in Ti*k*Z.

3 An overview of R

R is a language and environment for statistical computing and graphics. Its home page is `http://www.r-project.org`. R is a free software project. It compiles and runs on a wide variety of Unix platforms and similar systems, including FreeBSD, GNU/Linux,

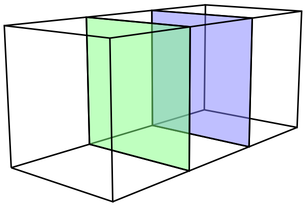

Figure 2: Graphic made in Sketch and exported to Ti*k*Z.

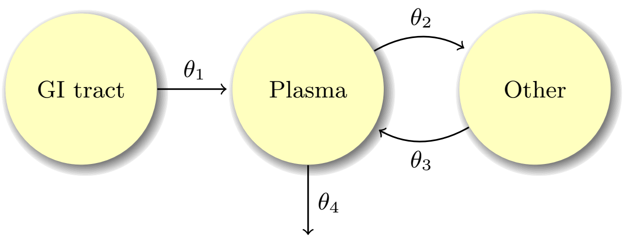

Figure 3: Hand-coded Ti*k*Z example.

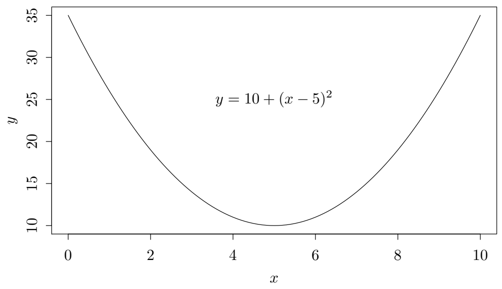

Figure 4: Example output using `tikzDevice`.

and Mac OS X, as well as Windows. R is often the vehicle of choice for research in statistical methodology, and it provides an open source route to participation in that activity.

R provides a wide variety of statistical techniques including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, and clustering. R is highly extensible and contains a rich collection of graphical techniques. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulas where needed. Defaults for the minor design choices in graphics have been carefully considered, but the user retains full control.

4 tikzDevice

The tikzDevice package enables LaTeX-ready output from R graphics functions. This is done by producing code that can be understood by the Ti*k*Z graphics language. All text in a graphic output with the `tikz()` function will be typeset by LaTeX and therefore will match whatever fonts are currently used in the document. This also means that LaTeX mathematics can be typeset directly into labels and annotations. Graphics produced this way can also be annotated with custom Ti*k*Z commands. An example R graphic output using tikzDevice is shown in fig. 4. The program that produced the graph is

```
setwd("~/tug2014")
source("quadratic-data.R")
source("quadratic-graph.R")
```

The `source` command executes the statements in the named file. Here I group code into small parts to focus discussion.

The file `quadratic-data.R` contains the data generation code:

```
x <- (0:100)/10
y <- 10 + (x-5)^2
```

R formulas are different from standard mathematical formulas. A function call operates on every element of a vector. When vectors of different lengths are added or subtracted, the shorter one is recycled in an effort to make them the same length.

The file `quadratic-graph.R` contains the graphics code:

```
require(tikzDevice)
tikz("quadratic-graph.tex",standAlone=FALSE,
                          width=4.5, height=2.5)
par(mex=0.6, mar=c(4.5,5,0,0)+0.1)
plot(x, y, type='l', xlab="$x$",ylab="$y$")
text(5, 25, "$y = 10+(x-5)^2$")
dev.off()
```

Most of this code is understandable by comparison to the resulting graph. An exception might be the `par` function used to change graphic parameters from their default values. The ones used here are:

Arg	Description
`mex`	A character size expansion factor used to describe coordinates in the margins of plots.
`mar`	Vector of the form c(bottom, left, top, right) which gives the number of lines of margin to be specified on the four sides.

5 Implementation

This section implements a dynamic document that facilitates reporting the current status of the race between Alison and Mitch. The document has a title, a graph of the posterior density, and a short statistical report. The data file for this senate race is `senate.dat` and contains just two numbers, the number in the sample that favor Alison, and the number that favor Mitch. This two-number file can be updated with an editor. For more complicated situations a program might update the data file.

Knitr is an R package containing a function `knit`, which takes a file name with an extension `.Rnw` as an argument. An `.Rnw` file is like a LaTeX file with interspersed R chunks. The output is a pure LaTeX file containing the output from running the R chunks. Documentation for `knitr` is available online and in Yihui Xie's book [1].

R is an implementation of a language S. There is a macro `\Sexpr()`, for *S expression*, that may be placed in the TEX portion of the file. `\Sexpr()` takes an R expression as an argument. The expression is evaluated, converted to text, and passed into the LATEX output. The content of `senate.Rnw` is

```
\documentclass[12pt]{article}
\begin{document}
<<setup,echo=FALSE>>=
source("chunk1.R")
@
\title{Alison Versus Mitch}
\author{David Allen\\University of Kentucky}
\maketitle
\thispagestyle{empty}
%
<<params,echo=FALSE>>=
source("chunk2.R")
@
A poll released July 28, 2014 produced
\Sexpr{a-2} potential voters favoring Alison and
\Sexpr{b-2} favoring Mitch.
These results give a \Sexpr{level} credible
interval for the proportion favoring Alison of
(\Sexpr{p1}, \Sexpr{p2}).
%
The posterior density function, with the area
over the credible interval shaded, is
<<label="density",dev='tikz',echo=FALSE,
  fig.width=4,fig.height=2.75,
  fig.align='center'>>=
source("chunk3.R")
@
\end{document}
```

The R chunks have a header of the form

```
<< ... >>=
```

where commands are inside the double brackets, on a single line (the line breaks above in the `<<label...` above are editorial). A chunk ends with an `@`.

The content of `chunk1.R` is

```
setwd("~/tug2014")
interval.length <- function(p1,a,b,level=0.95)
  {
  q <- qbeta(1-level, a,b)
  if( p1 > q ) return(1 - q)
  if( p1 < 0 ) return(qbeta(level, a, b))
  p2  <- qbeta(pbeta(p1, a, b) + level, a, b)
  return(p2-p1)
  }
```

The function `interval.length` is a function I wrote that gives the length of an interval starting at p_1. The posterior distribution of p is the beta distribution. `qbeta` is a built-in function giving quantiles of the beta distribution.

The content of `chunk2.R` is

```
vote  <- vector(mode="numeric")
vote  <- scan(file="senate.dat")
a     <- vote[1] + 2
b     <- vote[2] + 2
level <- 0.95
p1    <- optimize(f = interval.length,
          interval = c(0, qbeta(1-level, a,b)),
          a=a, b=b, level=level)$minimum
p2    <- qbeta(pbeta(p1, a, b) + level, a, b)
```

Here the built-in `optimize` function is used to find the value of p_1 associated with the shortest interval. Then the corresponding p_2 is computed.

The content of `chunk3.R` is

```
  left     <- (1:80)/80*p1
  interval <- p1 + (1:80)/80*(p2-p1)
  right    <- p2 + (1:80)/80*(1-p2)
  domain   <- c(left, interval, right)
  range    <- dbeta(domain, a, b)
  par(mex=0.6, mar=c(4.5,5,0,0)+0.1)
  plot(c(0,1), c(0, max(range)), type="n",
    xlab="Proportion Favoring Alison",
    ylab="Density",yaxt='n')
  polygon(c(interval, p2, p1),
    c(dbeta(interval, a, b), 0, 0), col=27)
  lines(domain, range); lines(c(0,1),c(0,0))
  lines(c(0.5,0.5),c(0,max(range)))
```

This code produces the graph.

After each data update, run the following command in a terminal:

```
Rscript -e "library(knitr);knit('senate.Rnw')"
```

A LATEX file is produced that can be processed in the usual ways. `Rscript` is just the command line version of R. The `-e` option means the following are statements to be run, as opposed to a file containing statements.

I conclude with an exercise. A poll was released on July 28, 2014 (the first full day of the conference in Portland) showing 321 for Alison and 336 for Mitch. I invite you to prepare a new data file, `senate.dat`, containing

```
321 336
```

Then *knit* the `.Rnw` file and LATEX the resulting `senate.tex`.

References

[1] Yihui Xie. *Dynamic Documents with R and knitr*. Chapman & Hall/CRC Press, 2014. ISBN 978-1482203530.

◇ David Allen
University of Kentucky
david dot allen (at) uky dot edu