

Dynamic Documents

David Allen
University of Kentucky

July 30, 2014

Presented at TUG 2014

1 Introduction

A generic definition of a dynamic document from Wikipedia:

A living document or *dynamic document* is a document that is continually edited and updated. A simple example of a living document is an article in Wikipedia, an online encyclopedia that permits anyone to freely edit its articles, in contrast to “dead” or “static documents”, such as an article in a single edition of the *Encyclopedia Britannica*.

The Approach Here

The approach here is to use the tools *R*, *tikzDevice*, *knitr*, and \LaTeX to produce a document that automatically updates when data changes. I start with an example. The tools are discussed along the way.

2 The Kentucky Senate Race

On November 4, 2014, the Commonwealth of Kentucky will elect a United States Senator. This race has high national impact and is closely watched.

The Candidates



Alison



Mitch

Presenting Polling Results

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.

Credible Interval

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.

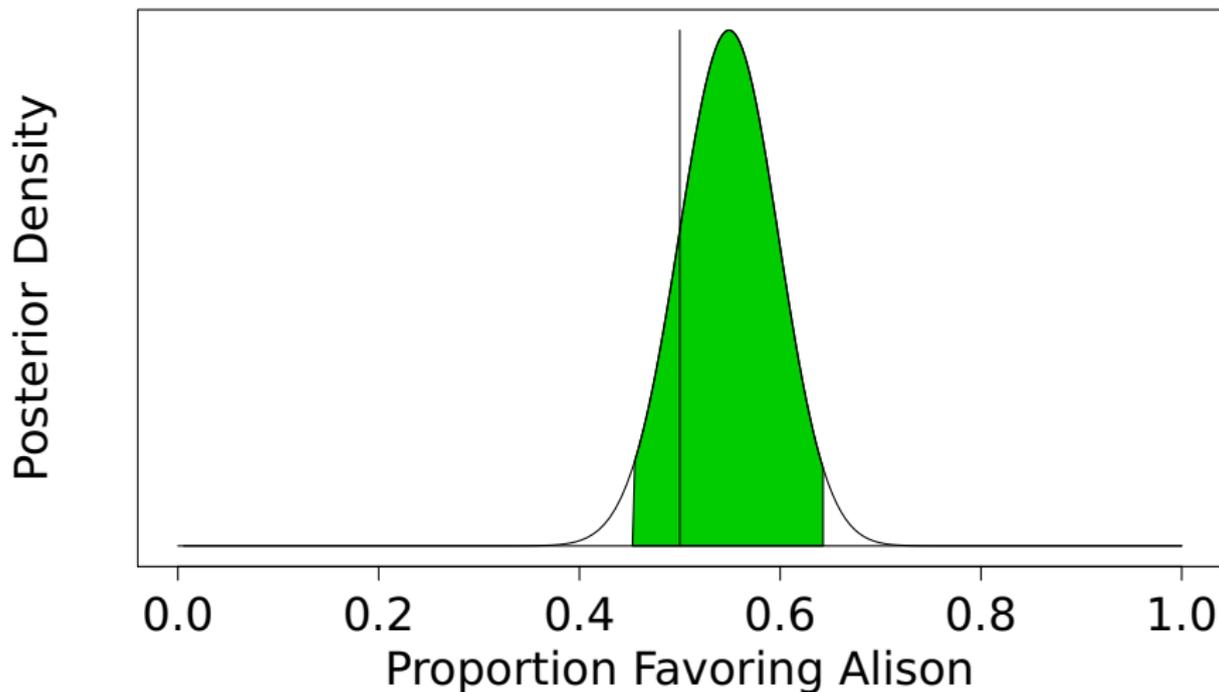
Details

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 $P(p_1 < p < p_2) = 0.95$. That probability statement does not uniquely determine the interval. The interval having minimal length is usually used.

An Example

An example assuming a sample with 55 favoring Alison and 45 favoring Mitch is shown on the next slide. The 0.95 credible interval (0.4528, 0.6428) is highlighted.

Posterior Density with Credible Interval



A “Report”

The cumulative results of polling through July 30, 2014 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).

Objective of this Presentation

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” on the preceding slide.

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. This presentation demonstrates the using knitr to automate the process.

3 *TikZ* graphics

TikZ is a graphics package used in conjunction with $\text{T}_\text{E}\text{X}$. It is included with most distributions of $\text{T}_\text{E}\text{X}$, but may be downloaded at

<http://sourceforge.net/projects/pgf/>.

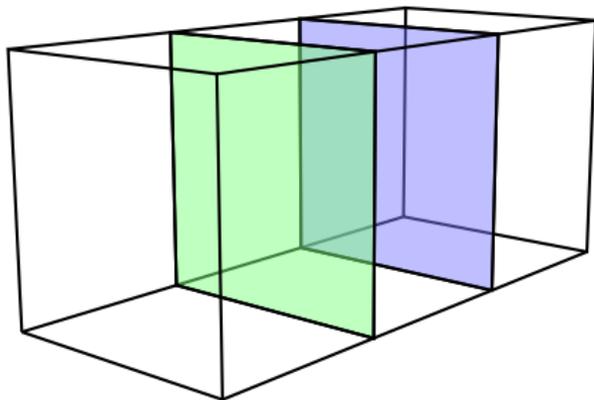
A large selection of examples of *TikZ* graphics are posted at

<http://www.texample.net/tikz/examples/>.

Examples I have composed are on the next two slides.

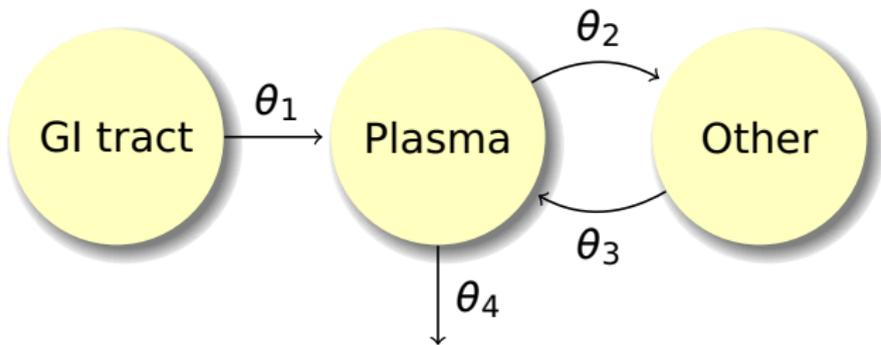
Fish Tank

This graphic was hand coded in Sketch,
<http://www.frontiernet.net/~eugene.ressler/>,
and then processed into *TikZ*.



A compartmental Model

This example was hand coded directly in *TikZ*.



4 A 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 and GNU/Linux), Windows and MacOSX. R is often the vehicle of choice for research in statistical methodology, and it provides an open source route to participation in that activity.

Statistical Procedures

R provides a wide variety of statistical techniques including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, and clustering.

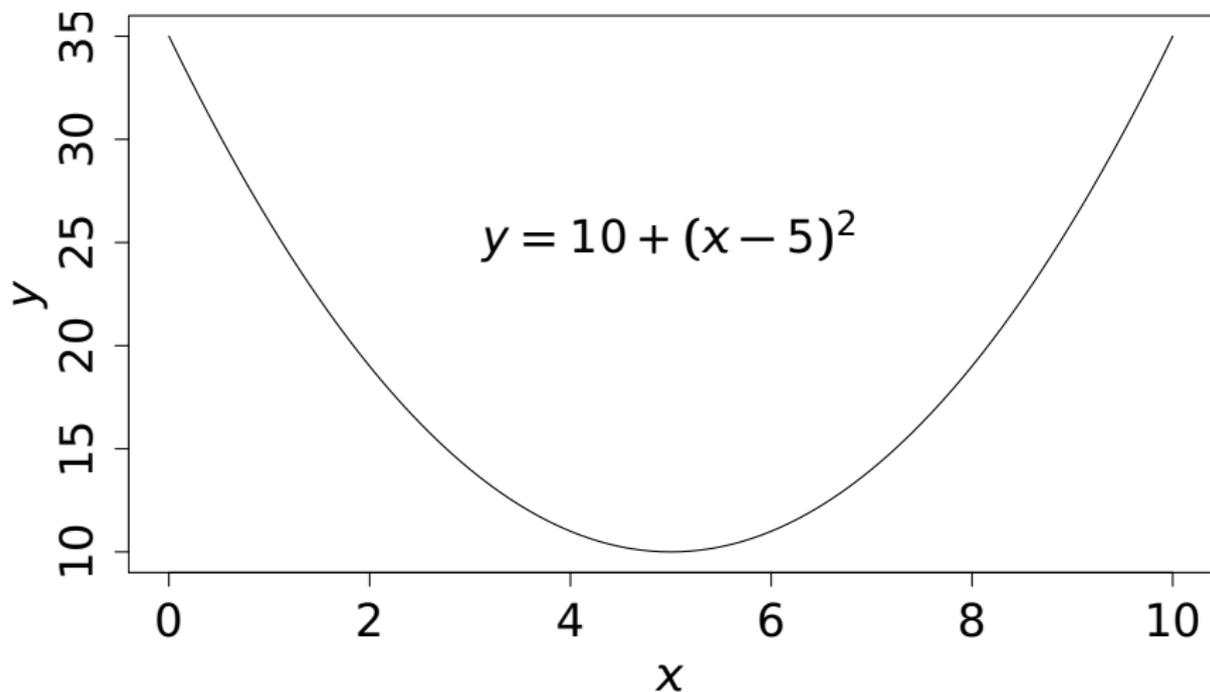
Graphics

R is highly extensible and contains extensive 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. Effort has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

5 A simple example

TikzDevice enables \LaTeX -ready output from *R* graphics functions. This is done by producing code that can be understood by the TikZ 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 TikZ commands.

The R Graphic using tikzDevice



An R program

The program that produced the preceding graph is

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

The “data”

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

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

Plotting code for graph

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()
```

Definition of selected par arguments

The “par” function is used to change graphic parameters from their default values. The ones used in this example are

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

6 Implementation

This section implements a dynamic document that facilitates reporting the current status of the race between Alison and Mitch. The document has

1. a title slide,
2. a graph of the posterior density, and
3. a short statistical report.

Data File

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. For this two person it can be updated with an editor. For more complicated situations there might be a program that updates a data file.

Knitr

Knitr is an *R* package containing a function `knit`. The function `knit` takes file name with an extension “.Rnw” as an argument. An .Rnw file is like a \LaTeX file with inter-dispersed *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].

Access to R variables

R is an implementation of a language S . There is a function $\backslash S\text{expr}()$, for S expression, that may be placed in the $\text{T}_{\text{E}}\text{X}$ portion of the file. $\backslash S\text{expr}()$ takes an R expression as an argument. The expression is evaluated, converted to text, and passed into the $\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$ output.

The senate.Rnw File

```
\documentclass[12pt]{article}
\usepackage{screen}
\begin{document}
<<setup,echo=FALSE>>=
source("chunk1.R")
@
\title{\color{TitleColor}
  Alison Versus Mitch}
\author{David Allen\\University of Kentucky}
\maketitle
\centerline{Presented at TUG 2014}
\thispagestyle{empty}
%
```

```
<<params,echo=FALSE>>=
```

```
source("chunk2.R")
```

```
@
```

```
\titlescreen{A ‘‘Report’’}
```

The cumulative results of polling through
\today\ 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}).

```
%
```

```
\titlescreen{Posterior Density Function}
```

```
<<label="density",dev='tikz',echo=FALSE,fig.width=
```

```
source("chunk3.R")
```

```
@
```

```
\end{document}
```

chunk1.R

```
setwd("~/tug2014/polling")
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)
}
```

chunk2.R

```
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)
```

chunk3.R

```
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)))
```

Processing senate.Rnw

After each data update run the following command in a terminal.

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

This produces a \LaTeX file that is processed in the usual ways.

Recent Polling Results

U.S. Senate Minority Leader Mitch McConnell has edged ahead of Democrat Alison Lundergan Grimes for the first time in a Bluegrass Poll, though the race for one of Kentucky's Senate seats remains a tossup.

With less than 100 days until Election Day, McConnell has taken a two-point lead over Grimes — 47 percent to 45 percent — as Republicans and coal-producing regions of Kentucky coalesce around McConnell, President Barack Obama's favorable rating remains low and McConnell appears to have neutralized the gender gap.

The poll of 714 registered voters was sponsored by the Herald-Leader and WKYT-TV in Lexington and The Courier-Journal and WHAS-TV in Louisville. It was

conducted by SurveyUSA from July 18 through July 23 and has a margin of error of plus or minus 3.7 percentage points.

Demonstration

The data file is

321 336

A script to process it, senate-report, is

```
#!/bin/sh
Rscript -e "library(knitr); knit('senate.Rnw')"
pdflatex senate.tex
evince --presentation senate.pdf
```

References

- [1] Yihui Xie. *Dynamic Documents with R and knitr*. Chapman & Hall/CRC Press, 2013. ISBN 978-1482203530.