Pergamon

Computers & Geosciences Vol. 23, No. 5, pp. 587–592, 1997

Printed in Great Britain
PII: S0098-3004(97)00027-7
0098-3004/97 $17.00 + 0.00

THE UCLA ELECTRONIC STATISTICS TEXTBOOK

JAN DE LEEUW
UCLA Statistics Program, 405 Hilgard Avenue, Los Angeles, CA 90095-1554
(*e-mail*: deleeuw@stat.ucla.edu)

(Received 23 September 1996; revised 10 November 1996)

Abstract—*UCLA Electronic Statistics Textbook* is an evolving bare-bones version of an electronic textbook. It brings together existing material that is appropriate to be displayed on the WWW, and tries to integrate this material into an integrated product. The emphasis is on hypertext, graphics, and interactive demos. © 1997 Elsevier Science Ltd

Key Words: World Wide Web, Education, Statistics.

INTRODUCTION

The UCLA Electronic Statistics Textbook[1]*, from now on UCLA-EST, is an attempt to write a statistics textbook which is:

1. freely available to everyone on the Internet;
2. independent of the level of the student, that is, useful at the undergraduate, graduate, and postdoctoral level;
3. interactive, using graphics and demonstrations;
4. complete, that is, it covers most of statistical theory as traditionally taught.

We use the model of a classical statistics textbook to bring together hypertext describing the traditional concepts and techniques of statistics and graphical components illustrating these concepts and techniques.

This particular mix, except for the hypertext, would just be a classical textbook that can be read online; but we also add interactive graphics, demos, and calculators to the mix, thus creating something that combines a classical textbook with computer software for demonstrations. Ultimately, we also plan to add exercises, examples, glossaries, and instructors' manuals.

UCLA-EST is under construction, and it will probably be under construction forever. It is not a product, but a project. One of the major advantages of the electronic form is that there is no limit on the size of the book, and that material can be added at any time. Moreover, the number of printings and the number of editions are unlimited.

In this paper, we will give an overview of UCLA-EST, its relationships with some other projects, its structure, and some of the technical problems in its implementation. We will not include graphical representations of the pages. It is much better if the reader browses the textbook while reading this paper. Thus, we suggest you now go to:

http://www.stat.ucla.edu/textbook/[1].

*Numbers in square brackets are citations to list of undated internet references.

STRUCTURE OF UCLA-EST

Structure of a page

Each page of the textbook has a header consisting of the textbook logo (a colorful histogram), plus the title of the book, which is Statistics. The Study of Stability in Variation

Each page has a footer, which consists of text-based buttons linking to the glossary, the table of contents, the toolbox, the textbook background, the formulas section, the textbook homepage, and the UCLA statistics home page. Readers can go to these places from any textbook page.

The footer also has a contact address, a counter showing the number of visitors, the full URL of the page, and the date it was last modified. The counter is text-based, and is implemented with a CGI script, and the modification date comes from a server-side "include".

The implementation of headers and footers is explained in more detail in a later section on technical considerations.

Table of contents

The table of contents of the textbook is fairly traditional. The first two levels are as follows:

1. Introduction
    (a) Variables
    (b) Variation
    (c) Sampling

(d) Models
(e) Inference

2. Analysis of a Single Variable
   (a) A Single Variable, Descriptive.
   (b) A Single Variable, Several Conditions.
   (c) A Single Variable, in Time.
   (d) A Single Variable, with Structured Conditions

3. Analysis of a Pair of Variables
   (a) Cross Tables
   (b) Correlation
   (c) Probability Models for Bivariation

4. Analysis of Multivariables
   (a) Describing Multivariables
   (b) Decomposing Multivariables
   (c) Path Analysis, Factor Analysis, and friends
   (d) Probability Models for Multivariables
   (e) Eigenvalues and Friends

If your browser is able to read Javascript, then you get a better idea by browsing:

http://www.stat.ucla.edu/textbook/textbookol.html.

Again, many of the sections are empty, or almost empty, and more sections will be added.

PHILOSOPHY

There are undoubtedly those who think that an author's or editor's philosophy should not influence the contents of the textbook that they write or edit. I do not agree. Ultimately, one of the factors that will hold the textbook together is a strong influence of a certain interpretation of statistics. One should not forget that one of the major disadvantages of hypertext and on-line books is that they have, by definition, much less unity and cohesiveness as a bound book. They are collections of pages, and it is imperative to remind the reader at many points that they are still reading or browsing the same object.

Models and techniques

Statistics is about the properties of statistical techniques. Statistical techniques are tools to transform data (input) into representation (output)—with the aim of a gain in clarity, communicability, stability, and so on.

Statistical techniques can be studied in terms of their stability under small variations of the input, and in terms of their performance in idealized situations, which are often called models.

In this interpretation of statistics, the emphasis is strongly on the techniques. Models can be used to generate techniques, using principles such as least squares or maximum likelihood, but this is just auxiliary. After this initial step is carried out, the technique still has to be evaluated, and not just on the model that was used to generate the technique; that would be cheating.

Thus, in this interpretation, we do not care if models are true. It is not even clear what this implies—the question is whether models are useful, and they are if they generate stable and illuminating techniques.

Variables

It is useful to develop statistics using the general concept of a variable. This is discussed in the short first chapter of UCLA-EST, and also in some interesting publications by Donald Macnaughton[2].

Inference

The whole notion of statistical inference is murky. Of course, we want to make statements about the population from the sample. We want to interpolate and extrapolate, as in all of science, and as in many other activities. The problem begins if we want to justify this activity in rational or even logical terms. This has nothing to do with statistics.

We construct techniques, and we study their behaviour in idealized situations. We then apply the same techniques in real situations and we hope they work—which is a leap of faith. In some instances, we can persuade our peers, or the public at large, or the jury, that our technique has worked and is better than the competition. In other instances, we can only convince a small group of our colleagues who are working on the same narrow topic.

RELATED PROJECTS

Other UCLA projects

The Journal of Statistical Software (JSS). The Journal of Statistical Software[3] is an electronic journal published by UCLA Statistics. It has two main functions. In the first place, it provides a service to the statistical community, because it provides useful statistical software in a central location. Here, "useful" implies both relevant, well-tested and well-documented. All three aspects are guaranteed by the review process.

The central location aspect is particularly important, because the Royal Statistical Society has stopped publishing algorithms in Applied Statistics. Of course there is a large volume of statistical software in statlib[3], but it is poorly organized and, in many instances, sparsely documented.

The second function of the journal is to provide a "respectable" outlet for authors of software and manuals, so that they can use these publications in their academic record. Again, the peer review process is critical here. The relation of JSS with UCLA-EST is clear. In the textbook, we provide, among other things, a repository of instructional

software for statistics. The code is usually available, or will be made available, but the emphasis is on using the programs over the net. In JSS, we have a journal-type organization, i.e. a linear list of contributions ordered in time. Each unit is completely self-contained, except perhaps by the classical mechanism of references. In UCLA-EST, the organization is by topic, and hypertext takes care of the many links between different units.

The UCLA Xlisp-Stat Archive. Xlisp-Stat[4] is a statistical and graphics environment based on the Lisp language. It can be compared with S-plus, but in contrast, with S-plus, it is completely open and free. Xlisp-Stat also has better memory management, better dynamic graphics, and a much better extension language. However, it is a less complete and a less smooth product than S-plus, with less support.

UCLA Statistics maintains an archive of contributed software in Xlisp-Stat at www.stat.ucla.edu/archive/xlisp-stat. UCLA-EST uses many interactive graphical demos written in Xlisp-Stat. Otherwise, the differences are clear. Although the Xlisp-Stat Archive is organized by topic, the units are not connected, and often documented poorly. There is a subdirectory of the archive that deals specifically with statistics teaching (statistics/introstat in the archive), and this contains some well-developed environments, but without the hypertext.

The UCLA course pages. At www.stat.ucla.edu/courses, there is a menu of statistics courses at UCLA. In many of them, the instructors maintain a set of home pages with lecture notes, outlines, homework, chat rooms, case studies, and so on. Several of the components are linked into portions of UCLA-EST, which shows another way in which the textbook can be used. For a particular course, the instructor uses only particular pieces by providing the links to the students.

There are also many courses outside UCLA that link parts of the textbook.

Other EST projects

The Chance Project. The Chance Project[5], by Laurie Snell at Dartmouth, maintains the Chance Database. The Chance Database contains material useful for Chance Courses, which are quantitative literacy courses based on current chance events in the news. Clearly, the emphasis is on the case-studies, and the Chance Database is a repository of case studies.

It is clear that UCLA-EST can use materials from the Chance Data-base, because ultimately, case-studies and examples will have to be integrated in the textbook. Unfortunately, the news is not a stable entity. A component explicitly defined in terms of the news must be upgraded continuously, which is not desirable, even in an on-line textbook.

Hyper Stat. Some time ago, David Lane pioneered electronic statistics hypertext by publishing HyperStat, written in Applc's Hyper-Card. This is now being converted to HTML[6].

HyperStat is similar to UCLA-EST, in the sense that it is clearly a textbook. It is at an elementary level, and it is written to fit on a floppy disk. The emphasis is on the hypertext and the static graphics: there are no interactive calculators and demos.

GASP. The Globally Accessible Statistical Procedures[7] project originates with Webster West of the University of South Carolina. Its purpose is to make statistical procedures widely available over the Net. There are basically two classes of procedures, which we call "calculators" and "demos" in UCLA-EST. In a calculator, you enter input, often in a form, or by giving the program the URL of a dataset. You maybe also set some options. The program then returns the results in the browser window, or maybe over e-mail. A demo is usually designed for instruction—it illustrates statistical concepts graphically and often dynamically. GASP has CGI-based calculators and Java demos. They are simply listed, and not organized in any way (except for the distinction discussed previously).

Journal of Statistics Education (JSE). The Journal of Statistics Education[8] is an electronic journal, published by the statistics department at [North Carolina State University (NCSU)]. It publishes papers on statistics education, in many instances with downloadable software. Papers are written in HTML. JSE is closer to classical journals than JSS, because it works with volumes, and it is paper-oriented and not software oriented. JSS is a software archive in disguise, JSE is a journal in disguise. Obviously, material in JSE has been used as background for UCLA-EST, but nothing is used directly.

Statlib. Statlib is an archive of many things that have to do with statistics. There is some attempt at organization, by bringing material under a large number of headings, essentially making it into a number of archives. The amount of useful material in statlib is large, but searching and browsing are difficult unless you know precisely what you are looking for. Much of the software is poorly documented, and there is no subclassification in terms of statistical topics or statistical activities (teaching, consulting, research).

TECHNICAL CONSIDERATIONS

HTML versions

HTML has been changing rapidly over the last 2 years. Fortunately, there is now some order in the HTML world. One reason is the rivalry between Netscape and Microsoft. If Netscape adds a feature to its browser, such as an HTML extension, then Microsoft soon will add the same feature. Since

their browsers control 90% of the market, this guarantees a de-facto standard.

The actual standard, as usual, is much slower to develop. HTML 2.0[9] is generally accepted, but it did not incorporate many of the Netscape extensions. HTML 3.0 was a promising development, but it attempted too much. Thus, it was abandoned. It seems that HTML 3.2[10] is a reasonable standard, and we try to conform to this as far as possible.

Graphics

The common denominator for graphics on the WWW is still the gif format. Although sometimes jpeg graphics are used in UCLA-EST, if we construct graphics files, we use gif. The basic tools on UNIX are xfig[11], a drawing program that can save its output as gif files, and xv, the graphics viewer and manipulator[12]. Many other useful tools are available, of course, on the Macintosh.

For the CGI programs that have to produce gif files, we rely on Thomas Boutell's gd library[13], which can produce its gifs "on-the-fly". Thus, they need not be stored anywhere on the system, and we merely use tags such as:

```
<IMG SRC = ''/cgi-bin/foo.cgi''>
```

where foo.cgi is the program generating the gif.

Now that "plug-ins" have become generally available (and portable over the main browsers), we can also consider using QuickTime movies. UCLA-EST has some QuickTime demos, written by Berrie Zielman (Netherlands).

Client-side versus Server-side

If we start a program from the HTML page, then in some situations, it is executed on the server (i.e. our machine), and in other situations on the client (i.e. your machine). Both options have advantages and disadvantages. If the program is executed on the server, then we need not assume anything about the client, and we can set things up so that they work correctly. However, a large number of users will cause an excessive load on our machine; and since the results must be sent over the network to the client, the response may be slow. On the other hand, if the program is executed on your machine, it must be downloaded first (in the example of a Java applet), and we are powerless to help you if your setup is not correct. Similarly, if your browser is not set up to deal with Java, JavaScript, client-pull, server-push, and so on, there is no way we can help.

Some aspects of this same problem will be discussed again in subsequent sections.

Java and Javascript

The three most popular browsers at the moment are Netscape's Navigator, Microsoft's Internet Explorer, and Apple's Cyberdog. All three understand Java, and the first two understand JavaScript (Cyberdog will understand JavaScript soon, because it will use a version of Navigator as its WWW browser). Also, all three browsers can be extended by using plug-ins (which make more file formats available for display in the main browser window).

Thus, increasingly, it seems that we can assume that Java and Javascript are available to our clients. UCLA-EST does not yet make this assumption, and warns the reader in the text that some demos will only work on some browsers. On the other hand, many excellent applets are now available, written by Balasubraminian Narasimhan (Stanford), Webster West (South Carolina), David Lane (Rice), Tony Rossini (South Carolina), and others. If they are available on the net, they are incorporated in the textbook in the appropriate section.

Helpers

UCLA-EST, in its current form, relies heavily on Xlisp-Stat demos. This implies that the person reading the textbook needs to have Xlisp-Stat installed correctly on their local machines. At the moment, we could run Xlisp-Stat on the server side, and have it display on the client machine, but this is a tricky process limited to clients running X11. We can also run Xlisp-Stat on the server side and have it display its text output in the browser window. This is actually done in

http://www.stat.ucla.edu/cgi-bin/Xlisp-Stat.cgi.

Unfortunately, we cannot do the same thing with graphical output, and only the client solution is open to us at the moment. One way around the dilemma is to rewrite the demos in Java, but a superior solution would be if Xlisp-Stat could write graphic output to a browser window.

CGI

UCLA-EST uses many CGI programs, mostly calculators. These are all written in C, using the cgic[14] library of Thomas Boutell. If the programs produce graphics, they use Boutell's gd[13] library, although most people think that CGI programs should be written in perl, and only masochists use C. The most interesting CGI programs in the textbook are perhaps the probability distribution calculators in

http://www.stat.ucla.edu/textbook/singles/describe_single/probmodels/calc.html.

All CGI processing takes place on the server-side, of course, which is one reason to use C instead of perl. Of course, everything that can be done using CGI could also be done with Java and JavaScript, at the expense of some portability. Observe that combining CGI and HTML forms even gives reasonable results in text-only browsers, such as lynx.

Preprocessing

UCLA-EST has hundreds of HTML pages, even in its current incomplete form. It is important, both

from the aesthetic point of view, and from the point of view that we want to emphasize the unity of the project, to make these pages look about the same. This is achieved by giving them identical headers and footers.

For the headers and footers, we use the pphtml program of[15]. This preprocesses a file foo.phtml of the form

```
<!-%/bin/sh TB_HEAD ''A Title''->
-insert HTML here-
<!-%/bin/sh TB_TAIL->
```

After processing, there will be a file foo.html in which the HTML comments have been replaced by the UCLA-EST header and footer. For completeness, we give TB_HEAD and TB_TAIL here. They are

```
echo '' <HMTL>''
echo '' <HEAD>''
echo '' <TITLE>$1 </TITLE>''
echo '' </HEAD>''
echo '' <BODY BGCOLOR = ''#FFFFFF''>''
echo '' <TABLE>''
echo '' <TR>''
echo '' <TD>''
echo '' <IMG SRC = \''/graphics/stat.gif\''ALIGN-LEFT>''
echo '' <TD VALIGN = BOTTOM> <FONT SIZE = 6>Statistics </FONT>BR>''
echo '' <FONT SIZE = 5>The Study of Stability in Variation </FONT>''
echo '' </TABLE>''
echo '' <HR>''
```

and

```
echo " < HR >"
echo '' <Center>''
echo ''[<A HREF = \''http://www.cas.lancs.ac.uk/glossary_v1.1/
main.html\''>''
echo ''Statistics Glossary </A>]''
echo ''[<A HREF = \''/textbook/textbookol.html\''>Table of Contents </
A>]''
echo ''[<A HREF = \''/calculators/\''>Statistics Toolbox </A>] <BR>''
echo ''[<A HREF = \''/textbook/background.html\''>Textbook Background </
A>]''
echo ''[<A HREF = \''/textbook/formulas/\''>Formulas </A>] <BR>''
echo ''[<A HREF = \''/textbook/\''>Textbook Homepage </A>]''
echo ''[<A HREF = \''/\''>UCLA Statistics Homepage </A>]''
echo '' </CENTER>''
echo '' <HR>''
echo '' <ADDRESS>''
echo ''Textbook Editor: Jan de Leeuw < BR>''
echo ''Email us at: ''
echo '' <A
HREF = \''mailto:deleeuw@stat.ucla.edu\''>deleeuw@stat.ucla.edu </A>''
echo ''Document: http://www.stat.ucla.edu < !-#echo
var = \''DOCUMENT_URI\''- < BR>''
echo ''Visitors since!-#exec cgi = \''/cgi-bin/counter-date\''-> < BR>''
echo -n ''Last revision: ''
date
echo '' </ADDRESS>''
echo '' </BODY>''
echo '' </HTML>''
```

Equations

Incorporating equations into HTML pages is problematic. There are many possibilities, but no perfect solution as yet. One long-awaited possibility is the < MATH > / < MATH > tags in HTML 3.0, which allowed for LaTeX-like commands directly in the HTML. However, these tags were implemented only in experimental browsers such as Arena[16], and the HTML 3.0 standard was abandoned recently.

A second possibility is to render your pages to either Postscript or pdf, and to display the page with a plug-in (for pdf) or helper (for Postscript). This assumes that the client is set up correctly to deal with these formats, and in the situation of Postscript, it removes all hypertext features. We can use hypertext links in pdf, but this requires the use of commercial software.

There are basically three ways to introduce equations into HTML documents in the common browsers. The first one is to use Java applets. This is achieved by EqnViewer[17] and by WebEQ[18]. Secondly, we can use a CGI program to preprocess the HTML page and render the equations. The prime example here is MINSE[19]. Finally, we can preprocess the equations to gif files, and incorpor-

ate them into the HTML. This is the most portable procedure, and it is the one used in UCLA-EST (although in future, we might switch to Java).

One brute force way to deal with equations is to write a document in LaTeX, and then use the latextohtml translator[20]. The translator translates the document into a number of linked HTML pages, and the equations are translated into gifs. Although this works well for some manuscripts, in other instances, it produces chaos. We prefer to have more control. For this control, once again we use pphtml.

We can preprocess phtml pages that contain HTML comments of the form

```
<!--%TEX
\begin{multline*}
\huge
\begin{bmatrix}
A & B
B' & C
\end{bmatrix}^{-1}=
\end{multline*}
-->
```

Here, TEX is a shell script, which contains

```
#!/bin/sh
DOC = $PPIDOC
NUM = $PPINUM
cat> $DOC-$NUM.tex < HEAD
\documentclass[12pt]{amsart}
\usepackage{amssymb}
\usepackage}uclastat,verbatim,float}
\unitlength 0.2 in
\thispagestyle{empty}
\begin{document}
\noindent
HEAD
cat >> $DOC-$NUM.tex
cat >> $DOC-$NUM.tex < FOOT
\end{document}
FOOT
latex $DOC-$NUM.tex >.latex-errors>
dvips $DOC-$NUM.dvi > $DOC-$NUM.ps
pstogif -out $DOC-$NUM.GIF $DOC-$NUM.ps >.pstogif-errors
giftrans -t 1 -b 0 $DOC-$NUM.GIF > $DOC-$NUM.gif
rm $DOC-$NUM.tex $DOC-$NUM.dvi $DOC-$NUM.ps
rm $DOC-$NUM.aux $DOC-$NUM.log $DOC-$NUM.GIF
echo '' < IMG SRC = \''$DOC-$NUM.gif\'' ALT = \''$DOC-$NUM\''>''
```

After preprocessing the file foo.phtml, the comments will have been replaced by<IMG SRC = ''foo-1.gif'' ALT = ''foo-1''>and the file foo-1.gif will be in the same directory as foo.html.

INTERNET REFERENCES

[1] UCLA Electronic Statistics Textbook: http://www.stat.ucla.edu/textbook/.

[2] Macnaughton, D. The Entity–Property-Relationship Approach to Statistics: An Introduction for Students: http://www.hookup.net/ donmac/..

[3] Journal of Statistical Software: http://www.stat.ucla.edu/journals/jss/..

[4] Tierney, L. Lisp-Stat Information: http://stat.umn.edu/~luke/xls/xlsinfo/xlsinfo.html..

[5] Chance Database Welcome Page: http://www.geom.umn.edu/docs/snell/chance/welcome.html.

[6] Lane, D. Overview of HyperStat: http://www.ruf.rice.edu/~lane/hyperstat/overview.html..

[7] Globally Accessible Statistical Procedures: http://www.stat.sc.edu/rsrch/gasp/..

[8] Journal of Statistics Education: http://www2.ncsu.edu/ncsu/pams/stat/info/jse/homepage/html..

[9] HTML 2.0 Proposed Standard Materials: http://www.w3.org/pub/WWW/MarkUp/html-spec/..

[10] Raggett, D. HTML 3.2 reference Specification: http://www.w3.org/pub/WWW/TR/WD-htm132.html..

[11] Xfig 3.1.4 Release Notes: http://www.madness.net/~vince/software/SGI_freeware_CD/relnotes/xFig.html..

[12] Bradley, J. Note on XV: http://www.sun.com/sunsoft/catlink/xv/note.html.

[13] Boutell, T. gd 1.2. A graphics library for fast GIF creation: http://www.boutell.com/gd/.

[14] Boutell, T. cgic: an ANSI C library for CGI Programming: http://www.boutell.com/cgic/.

[15] Thrift, P. Preprocessing instructions: Embedding external notations in HTML: http://www.ncsa.uiuc.edu/SDG/IT94/Proceeding/DDay/thrift/PPI.html..

[16] Arena: http://www.stat.ucla.edu/textbook/www.w3.org/pub/WWW/Arena/.

[17] Eqn Viewer on the Web: http://www.hookup.net/~cbazza/EqnViewer.html..

[18] WebEQ Equation Rendering: http://www.geom.umn.edu/software/WebEQ/..

[19] Ka-Ping Yee, Mathematics has arrived on the Web at last...: http://www.lfw.org/ping/..

[20] Nikos Drakos, All About LaTeX2HTML: http://cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html..

[3] Statlib: http://lib.stat.cmu.edu/..