Preparing Students for the Future of Statistics

Jay Emerson
August 9, 2015
KEEP CALM AND SAY IT DEPENDS
It Depends

- Who are the students?
- What is the objective of the course?
- What and how are you comfortable teaching?
  - Introductory? Undergraduate? Graduate? Professional?
  - Development of theory
  - Survey of methodology
  - Data analysis? What is a course in data analysis?
    - Answer: It depends.
Courses called "Data Analysis" often:

- Present a toolkit of methods
- Show "how to do (insert method here) in (insert statistical software here) with (this example or toy data set)"
- Use clean, processed, convenient data
- Belong in the "survey of methodology" category
Data Analysis

- The starting point is the data set – not the methodology
  - What do I have? Where did it come from?
  - Do I understand and trust these data enough to justify further analysis?
  - Are there errors?
  - Is cleaning or significant reorganization required?
- You can't spend too much time on these issues.
- You can't *prove* there are no problems with the data.
  - You have to look for them.
  - Not looking for problems makes data analysis an act of faith than of science.
Know your audience

- Particularly important in data analysis!
  - Range of backgrounds in mathematics, probability, or statistics.
  - Range of backgrounds in computer programming (in general) or with your statistical software (in particular).
    - Less important in a survey of methodology
    - More important in programming-intensive data analysis
- Be clear about your course objectives and prerequisites.
Course objective? Students?

Answer: it depends!
Diversity of student backgrounds

<table>
<thead>
<tr>
<th>Statistics, Programming</th>
<th>No Statistics, No Programming</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Statistics, Programming</td>
<td>No Statistics, Programming</td>
</tr>
</tbody>
</table>

Even this oversimplifies greatly. "Mathematical/QR maturity"?

- I'll join Susan in focusing on the top row.
- I think Dick is about address the bottom row.
Consider a non-introductory course and resist the temptation to teach lots of "new stuff". Instead:

- Re-teach as necessary.
- Deepen understanding of important topics.
- Push programming/scripting and problem solving skills to another level.
- **Primary Objective: "distance traveled".**
- Don't use a book.
- Small-ish class? Take the time to stick your nose in your students' code, while sitting next to them.
- Large-ish class? Make sure you have great Teaching Assistants.
Survey of methodology "data analysis"

data(iris)
head(iris)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa

km.iris <- kmeans(iris[,-5], centers=3, nstart=10)
Real data analysis

Even real examples you've studied before aren't "easy" to recycle – in the real world stuff always goes wrong!

```r
page <- 'http://www.goldsheet.com/gs_new/historic/cbblog12.html'
x <- scan(page, what='', sep='
')
head(x)
```

```r
## [1] "<div style="text-align: center; font-size: 14px; font-weight: bold">"
## [2] "    <p>You have reached the newly upgraded Goldsheet.com. The page you are trying to reach is no longer part of our site.<p>
## [3] "    <p>To view our site please follow this link to <a href="http://goldsheet.com">http://goldsheet.com</a> where you can review all our new navigation options.<p>
## [4] "    <br />
## [5] "    Thank you!<br />
## [6] "    - Goldsheet<br />
```

11/30
Real data analysis, second attempt:

```r
x <- scan("cbblog12.html", what="", sep="\n")  # They moved it, I got it.
x[1000]
```

```r
```

```r
x <- gsub(" &nbsp;", "", x, fixed=TRUE)  # HTML space codes
x <- gsub(" </span>" , ",", x, fixed=TRUE)  # Make use of HTML structure...
x <- gsub(" &lt;[^<>]*&gt;" , ",", x)  # ... then get rid of the HTML
x[1000]
```

```r
```

12/30
Problem-solving

- Is programming/scripting important?
- My answer: yes! In my world, the answer is not "it depends".
  - An important quantitative reasoning (QR) skill, and
  - QR is an important part of a liberal arts education.
- So I'd challenge you to think about…
Programming/scripting for serious, flexible problem-solving
Specialized tools: useful but inflexible
Closing advice on data analysis courses, for your students, and for you

I'll then close, time permitting, with:

- Something for you to argue about over beer tonight
- Thanks
- Something for you to look up if it interests you
Real data analysis courses

- Real-world (often messy) data: can be hugely time-consuming.
- More difficult "prep" than a more conventional theory or methodology course.
- The course will (and should) change every time you teach it.
- Don't be afraid to say "I don't know" or "I'm not sure how to solve that problem" – then follow it up with "let's figure it out."
- Be ready to say "It depends."
For your grad students or advanced undergrad majors:

- Become at least an intermediate programmer/scripter in a high-level statistical software environment. **Probably R. Possibly Matlab/Julia/Python/…** (but see below). It depends.
- Have at least some basic ability to:
  - Program/script in a language like Python (or Perl, …)
  - Program/script in C (or C++)
For you

- See the previous slide.
- Practice. You don't learn to speak French by reading a book, even if the book may be a helpful resource.
- Build a foundation.
  - "Base R" (no packages at first) as a language rather than as a toolkit of statistical methodologies.
  - A very small subset of the R language supports 95% of my daily coding.
- It's never too late to start. But it won't get easier if you delay.
For debate over drinks tonight

One perspective expressed at a conference this summer:

- Data analysis should be scripted, not clicked.
- Reproducibility is important.
- Communication is best achieved via code.
- The overwhelming majority of bottlenecks in data analysis and statistical computing are human (not software or hardware).
For debate over drinks tonight

Surprisingly heated debate and dissent (roughly a 50-50 split):

- What is meant by "programming"?
- We don't need to turn everyone into computer scientists.
- A good GUI generates scripts... isn't that enough for reproducibility and communication requirements?
- Saying that "point-and-click statistical analysis is only for people who can't code" is a poor position to take.
Thanks

Much of my thinking about data analysis and data analysis courses has been heavily influenced by John Hartigan. Errors or omissions are my own fault.

Time permitting, an extended conclusion...
Abstract

The evolution of computing is currently in a period of rapid change. These changes are certain to provide major opportunities and challenges in statistics. We look at history and current trends, both in general computing and in statistical computing, with the goal of identifying key features and requirements for the near future.
Human Needs

We will emphasize human needs, both for the users of an interactive statistical system and for the developers of new or modified statistical software.
What is Labor Intensive?

The writing and testing of programs have become an increasingly large part of the cost of using computers.
Skilled Human Labor

The important measurement here is against the value of the scarcest resource in most environments: skilled human labor. The most important single implication is the need to consider the wise use of human effort more carefully than has often been the case.
Commentary

- The last four slides contained approximately 150 words.
- I wrote (or mildly edited) about 3 of them. I stole the rest.
- I deleted a few phrases for impact and brevity.
- The original text: by John Chambers (creator of the S programming language and R core member) in *The American Statistician* …

- published in **1980**!

- In some ways, not much has changed in 35 years.
A Tip of the Hat to John Chambers: The S Language and R Core Member

![Image of John Chambers]
Conclusion

The entire field of computing is in a period of rapid expansion and change. Many exciting challenges face those who want to extend our capabilities in the statistical use of computers. The advances in hardware and software make possible more effective and widespread computing for data analysis. The most valuable resource is skilled human effort; computing should be organized to improve the ease and the quality of our work, whether we are users of statistical systems or designers and implementers of software.
Conclusion from Chambers (1980)

The entire field of computing is in a period of rapid expansion and change. Many exciting challenges face those who want to extend our capabilities in the statistical use of computers. The advances in hardware and software make possible more effective and widespread computing for data analysis. The most valuable resource is skilled human effort; computing should be organized to improve the ease and the quality of our work, whether we are users of statistical systems or designers and implementers of software.