CSE 599F
Research Methods and Data Analysis
in Software Systems Research
Winter 2019

Data visualization
March 05, 2019

Recap

- Monte Carlo (MC) methods
  - What is the key concept?
  - When are these applicable and useful?
- Bootstrapping (in R)
  - What is the key concept?
  - How is bootstrapping different from and related to MC methods?
  - What is the core assumption for bootstrapping?

Today

- From an experimental design to a research paper
- Data visualization
  - Guidelines for effective tables and graphs
  - Effective tables with *booktabs*
  - Effective graphs with *ggplot2*

Empirical research: the pipeline

Experimental Design → How do we get here? → Research paper
Empirical research: the pipeline

Experimental Design
- Data collection
- Data analysis
- Graphs & tables

Research paper

Do all results go into the paper?

Sanity checks

Data collection
- Data analysis
- Graphs & tables

Research paper

Detailed results

Effective tables

Tables vs. graphs

**When are tables useful?**
- Compare individual values.
- Precise values are important.
- Values may involve multiple units.

**When are graphs useful?**
- Consider an entire set of values.
- Visualize trends and patterns.
- Relative differences and relationships are more important than precise values.

<table>
<thead>
<tr>
<th>Browser</th>
<th>Market share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet Explorer</td>
<td>78.4</td>
</tr>
<tr>
<td>Firefox</td>
<td>18.9</td>
</tr>
<tr>
<td>Safari</td>
<td>2.8</td>
</tr>
<tr>
<td>Chrome</td>
<td>—</td>
</tr>
<tr>
<td>Opera</td>
<td>2.1</td>
</tr>
<tr>
<td>Netcrape</td>
<td>0.5</td>
</tr>
<tr>
<td>Other</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Market share 2018
Effective tables: layout

A first table in LaTeX:
- Recall the run-time data set
- Goal: show run times and improvements


table

| variant, naïve, caching, forking, equivalent, first.order, run, subject |
|-------------------------|---------------------|--------|-----|------|------|--------|
| 11, 309.8, 157.6, 144.8, 1, 1, 1, “tax” |
| 12, 379.5, 237.4, 254.5, 0, 1, 1, “tax” |
| 13, 415.9, 225.9, 225.9, 0, 0, 1, “tax” |
...

Looks pretty good and clear, doesn’t it?

Effective tables: layout

A first table in LaTeX:
- Recall the run-time data set
- Goal: show run times and improvements

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>RUN TIMES AND IMPROVEMENTS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>RT-naive</td>
</tr>
<tr>
<td>tax</td>
<td>504.11</td>
</tr>
<tr>
<td>ticTacToe</td>
<td>17.44</td>
</tr>
<tr>
<td>triangle</td>
<td>3.13</td>
</tr>
</tbody>
</table>

Which table looks better?

Effective tables: content

Still room for improvements -- keep it simple:
- Avoid mixing higher-is-better and lower-is-better values.
- Allow for easy comparisons, primarily by row.
- Summarize the table (what is the bottom line?).

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>RUN TIMES AND IMPROVEMENTS FOR THE NAÏVE, CACHING (CACHE), AND FORKING (FORK) STRATEGIES. RUN TIMES ARE GIVEN IN SECONDS AND AVERAGED OVER FIVE RUNS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Run times (cache vs. naïve)</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Tax</td>
<td>504</td>
</tr>
<tr>
<td>TicTacToe</td>
<td>17.4</td>
</tr>
<tr>
<td>Triangle</td>
<td>3.13</td>
</tr>
</tbody>
</table>

Still room for improvements -- keep it simple:
- Avoid mixing higher-is-better and lower-is-better values.
- Allow for easy comparisons, primarily by row.
- Summarize the table (what is the bottom line?).
Effective tables: summaries

<table>
<thead>
<tr>
<th>Subject</th>
<th>LOC</th>
<th>Speed up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax</td>
<td>8900</td>
<td>10.2%</td>
</tr>
<tr>
<td>TicTacToe</td>
<td>120</td>
<td>54.2%</td>
</tr>
<tr>
<td>Triangle</td>
<td>80</td>
<td>60.9%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>3393</strong></td>
<td><strong>41.8%</strong></td>
</tr>
</tbody>
</table>

Total vs. Average

How to properly summarize a table?

Effective tables in LaTeX:
- Use the `booktabs` package!
- Use descriptive (hierarchical) headers.
- Make each table self-contained (content and caption).
- Don’t use horizontal lines between related rows.
- Don’t use vertical lines between related columns.
- Right align numbers.
- Summarize with meaningful totals or weighted averages.
- Think about precision vs. significant digits (be consistent)!

Effective graphs: taming complexity

Way too many details!
How can this plot be simplified and improved?
Effective graphs: axes

Truncated axes are misleading and not a proper way to “demonstrate” effect size.

Effective graphs: point plots vs. line plots

Point plots (scatter plots):
- Good visual summary of point clouds, trends, and relationships.
- May obscure relevant trends (overlapping points).
- Hard to reason about density (without adding transparency).

Effective graphs: histogram vs. kernel density

Histograms:
- Good visual summary of count data.
- Binning may lead to misleading results.
- Kernel density overlay can provide information about adequate binning.

Effective graphs: box plots vs. violin plots

Box plots:
- Good visual summary for continuous data.
- Nicely complements hypothesis tests.
- May be misleading for multimodal data.
Effective graphs: facet plots

Effective graphs: ggplot2 example

# Point plot of the raw data
```r
ggplot(data=result, aes(x=Step, y=Ratio, color=Strategy)) + geom_point() + xlab("Work") + ylab("Test completeness")
```

# Aggregated line plot of the same data
```r
ggplot(data=result, aes(x=Step, y=Ratio, color=Strategy)) + stat_summary(fun.y="mean", geom="line") + xlab("Work") + ylab("Test completeness")
```

Effective graphs

Effective graphs in R:
- Use the ggplot2 package!
- Make each plot self-contained (content and caption).
- Relate tables and graphs to tell a coherent story.
- Avoid multiple, unrelated axes.
- Consistently put the DV on the vertical axis (by default).
- Choose an appropriate graph for the data (don’t connect unrelated data points).
- Reduce complexity with facet plots.