What is statistical graphics?

VIGRE, Tuesday Sep 7, 2004
1. Outline

- Why does plotting data produce richer solutions to problems?
  - Tipping data
  - Textbook solution
  - What's different when we plot the data

- Why has the forefront of statistical graphics moved beyond static graphics?
  - Quality control of olive oils
  - Simple classification of regions
  - Difficult classification of southern areas

- How about a game?
2. Why plot data?

2.1. Tipping data: Textbook solution

One waiter records 244 dining parties for 2.5 months, early 1990. Recorded total tip, total bill, sex of payer, smoking or not, day of the week, time of day, size of the party.

What are the important factors in tipping behavior?


2.2. Textbook solution

\[ \hat{\text{Tip Rate}} = 0.1844 - 0.0092 \times \text{Size}, \quad R^2 = 0.02 \]

Have you seen the signs in restaurants about tip being included for larger parties?
2.3. What's different when we plot the data

What's the distribution of tips?
How does tip relate to total bill?

![Scatter plot showing the relationship between total bill and total tip, with a correlation coefficient of r=0.68.](image-url)
How do demographics change tip relation to total bill?

<table>
<thead>
<tr>
<th></th>
<th>Total Bill</th>
<th>Total Tip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male Non-smokers</strong></td>
<td>r=0.82</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Female Non-smokers</strong></td>
<td>r=0.83</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>Male Smokers</strong></td>
<td>r=0.48</td>
<td>0.48</td>
</tr>
<tr>
<td><strong>Female Smokers</strong></td>
<td>r=0.52</td>
<td>0.52</td>
</tr>
</tbody>
</table>
How do demographics change with time of day?
Summary

- There is a tendency to give “cheap” tips rather than “generous” tips relative to the total bill.

- There is a tendency to round the tip to the nearest dollar, and to a lesser extent to the nearest 50c.

- There is much more variation in the tips given by smoking parties than non-smoking parties. And the female non-smokers are very consistent tippers.
3. Beyond static graphics

3.1. Data problem: olive oils

How do we distinguish the oils from different geographic regions in Italy based on their fatty acid composition?

Percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) for 572 samples from 9 areas (3 regions) in Italy.

Reference: Forina et al. 1983, “Food Research and Data Analysis”. 
3.2. The three regions: a simple classification

Linear discriminant analysis solution:

<table>
<thead>
<tr>
<th>True/Pred</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>322</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>6</td>
<td>145</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LD1</th>
<th>LD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>palm</td>
<td>-0.003</td>
<td>0.008</td>
</tr>
<tr>
<td>p’oleic</td>
<td>-0.014</td>
<td>0.018</td>
</tr>
<tr>
<td>stearic</td>
<td>0.002</td>
<td>0.004</td>
</tr>
<tr>
<td>oleic</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>lino</td>
<td>-0.001</td>
<td>-0.002</td>
</tr>
<tr>
<td>l’enic</td>
<td>-0.036</td>
<td>0.005</td>
</tr>
<tr>
<td>arach</td>
<td>0.016</td>
<td>-0.038</td>
</tr>
<tr>
<td>eico</td>
<td>-0.160</td>
<td>0.012</td>
</tr>
</tbody>
</table>
Tree solution:

If eicosenoic >= 7 then assign the sample to the SOUTH (1).
Else
    If linoleic>=1054 assign the sample to the SARDINIA (2).
    Else assign the sample to the NORTH (3).

<table>
<thead>
<tr>
<th>True/Pred</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>323</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>151</td>
</tr>
</tbody>
</table>
Using manual tour controls:

Limitations of methods:
LDA: Assume groups have homogeneous variance-covariance.
Trees: Uses a single variable at each split.
3.3. Difficult classification of southern areas

Looking at the data, using tours, it looks like a messy problem.

Explanation: Sicily’s demand exceeds supply, and sometimes borrows olives from the neighboring areas.
Summary:

- The **oils** from growing areas are quite **distinct, except for Sicily**!
- **Simple rules are ok**: Southern oils are distinguishable by a presence of eicosenoic acid. North and Sardinia separated by linear combination of linoleic and arachidic acid. Southern areas separated by linear boundaries, excluding Sicily.
4. Summary

- **Myth 1:** Plotting your data biases the result.
  Plots provide richer insight into problems, and solutions.
- **Myth 2:** Plots are difficult to summarize.
  Plots can provide simpler explanations to problems.
- **Myth 3:** We can’t see beyond 3D.
  Ready to play games in 4D?
5. A game called “Face Off”

Goal: Be the first to claim a face of the 4D cube with your color