This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.
How to display data badly

Karl W Broman
Department of Biostatistics

http://www.biostat.jhsph.edu/~kbroman

Using Microsoft Excel to obscure your data and annoy your readers

Karl W Broman
Department of Biostatistics

http://www.biostat.jhsph.edu/~kbroman
Inspiration

This lecture was inspired by

American Statistician 38(2):137-147

Dr. Wainer was the first to elucidate the principles of the bad display of data.

The now widespread use of Microsoft Excel has resulted in remarkable advances in the field.

General principles

The aim of good data graphics:
Display data accurately and clearly.

Some rules for displaying data badly:
– Display as little information as possible.
– Obscure what you do show (with chart junk).
– Use pseudo-3d and color gratuitously.
– Make a pie chart (preferably in color and 3d).
– Use a poorly chosen scale.
– Ignore sig figs.
Example 1

Example 2

Distribution of genotypes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>21%</td>
</tr>
<tr>
<td>AB</td>
<td>48%</td>
</tr>
<tr>
<td>BB</td>
<td>22%</td>
</tr>
<tr>
<td>missing</td>
<td>9%</td>
</tr>
</tbody>
</table>
Example 3

Example 4
Example 5

Example 6
Example 7

<table>
<thead>
<tr>
<th>N</th>
<th>b/c = 10.0</th>
<th>b/c = 10.0</th>
<th>b/c = 100.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r^*$</td>
<td>$G$</td>
<td>$r^*$</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.20</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.26</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.32</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.38</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.45</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.51</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.57</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0.63</td>
<td>4</td>
</tr>
</tbody>
</table>

Example 8

[Graph showing proportions survived vs log dose for Drug A, Drug B, and Drug C]
Displaying data well

• Be accurate and clear.
• Let the data speak.
  – Show as much information as possible, taking care not to obscure the message.
• Science not sales.
  – Avoid unnecessary frills — esp. gratuitous 3d.
• In tables, every digit should be meaningful. Don't drop ending 0’s.

Further reading

• WS Cleveland (1994) The elements of graphing data. CRC Press.
Displaying distributions

Dotplots

- Few data points per group
- Possibly many groups
Histograms

- Many data points per group
- Few groups
- Area of rectangle prop. to number of data points in interval
- I like many bins (typically $2\sqrt{n}$)

Boxplots

- Many data points
- Possibly many groups
- Displays minimum, lower quartile, median, upper quartile and maximum
### Summary statistics

<table>
<thead>
<tr>
<th>Location/Center</th>
<th>mean (i.e., average)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median</td>
</tr>
<tr>
<td></td>
<td>mode</td>
</tr>
<tr>
<td></td>
<td>geometric mean</td>
</tr>
<tr>
<td></td>
<td>harmonic mean</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale</th>
<th>standard deviation (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>inter-quartile range (IQR)</td>
</tr>
<tr>
<td></td>
<td>range</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other</th>
<th>quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>quartile</td>
</tr>
<tr>
<td></td>
<td>quintile</td>
</tr>
</tbody>
</table>

### The means

\[
\text{mean} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{(x_1 + x_2 + \ldots + x_n)}{n}
\]

\[
\text{geometric mean} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp \left\{ \frac{1}{n} \sum_{i=1}^{n} \log x_i \right\}
\]

\[
\text{harmonic mean} = \frac{1}{\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right\}}
\]
Measures of location/center

- Forget about the **mode**.
- The **mean** is **sensitive** to outliers.
  (The balance point for the histogram.)
- The **median** is **resistant** to outliers.
- The **geometric mean** is used when a logarithmic transformation is appropriate (for example, when the distribution has a long right tail).
- The **harmonic mean** may be used when a reciprocal transformation is appropriate (very seldom).
A key point

The different possible measures of the “center” of the distribution are all **allowable**.

- Which is the best measure of the “typical” value (for your situation)?
- Be sure to make clear which “average” you use.
Standard deviation (SD)

Sample variance
\[ \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2 \]

Sample SD
\[ \sqrt{s^2} = s \]
= RMS (distance from average)
= "typical" distance from the average
= sort of like \[ \text{ave}\{|x_i - \bar{x}|\} \]

Note:
\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]
Skewed data

Mean

Median

SD