Creating effective figures and tables

Karl W Broman
Biostatistics & Medical Informatics
University of Wisconsin – Madison

kbroman.org
github.com/kbroman
@kwbrroman
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.
Show the data

[Graph showing data for Treatment and Control groups with response on the y-axis and group on the x-axis.]
Show the data
Show the data
Show the data
Show the data
Avoid pie charts
Avoid pie charts
Avoid pie charts
Avoid pie charts

via @MonaChalabi (bit.ly/pie_vs_barchart)
Avoid pie charts

via @MonaChalabi (bit.ly/pie_vs_barchart)
Avoid pie charts

via @MonaChalabi (bit.ly/pie_vs_barchart)
Avoid pie charts

Via @MonaChalabi (bit.ly/pie_vs_barchart)
Consider logs
Consider logs
Consider logs
Consider logs
Consider logs
Consider logs
Consider logs
Consider logs
Take differences

\[ \log_2 \{ \frac{R_1}{R_2} \} \]

\[ \text{Ave} \{ \log_2 R_1, \log_2 R_2 \} \]
Ease comparisons

(things to be compared should be adjacent)
Ease comparisons

(add a bit of color)
Which comparison is easiest?
Don’t distort the quantities
(value $\propto$ radius)

Wheat (17 Gbp)
Arabidopsis (0.145 Gbp)
Human (3.2 Gbp)
Don’t distort the quantities
(value $\propto$ area)

- Wheat (17 Gbp)
- Arabidopsis (0.145 Gbp)
- Human (3.2 Gbp)
Don’t use areas at all
(value \propto length)
Quantities

- Position
- Length
- Angle
- Area
- Luminance (light/dark)
- Chroma (amount of color)

Categories

- Shape
- Hue (which color)
- Texture
- Width
Ease comparisons

(align things vertically)
Ease comparisons

(use common axes)
Use labels not legends
Don’t sort alphabetically
Must you include 0?

<table>
<thead>
<tr>
<th>Method</th>
<th>Detection rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>96.5%</td>
</tr>
<tr>
<td>B</td>
<td>98.1%</td>
</tr>
<tr>
<td>C</td>
<td>99.2%</td>
</tr>
</tbody>
</table>

![Bar chart showing detection rates for methods A, B, and C.](chart.png)

![Line graph showing detection rates for methods A, B, and C.](graph.png)
### A bad table

<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>0.2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.26333</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.32333</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.38267</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.446</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.50743</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.56743</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0.62948</td>
<td>4</td>
</tr>
</tbody>
</table>
### Fewer digits

<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>
In the past two decades globally, noticeable increases in the absolute numbers of people with stroke took place in the absolute numbers of people with stroke in low-income and middle-income countries. Presently, age-standardised rates of stroke mortality in people aged 75 years and older (table 1) were almost 30% more survivors younger than 75 years and 21% in people aged 75 years and older (table 1). Conversely, the number of stroke survivors (113%), DALYs lost (31%), and stroke-related deaths (26%), and DALYs lost (12%; 20% [15–30] in low-income and middle-income countries) were in people aged 75 years and older (table 1). The most striking increases in the number of strokes related deaths (26%), and DALYs lost (12%; <75 years: 36% [24% in high-income countries and 20% [15–30] in low-income income level (a 12% [95% CI 6%–17%] statistically significant 25% reduction in mortality rate (37% [31–41] in high-income countries and 20% [15–30] in low-income and middle-income countries). Further more, there was a significant 27% (19–43) in high-income countries and 22% [18–32], respectively), and mortality-to-incidence ratio (MIR) by age groups in high-income and low-income and middle-income countries, and globally in 1990, 2005, and 2010

### Table 1: Age-adjusted annual incidence and mortality rates (per 100 000 person-years), disability-adjusted life-years (DALYs) lost, prevalence (per 100 000 people), and mortality-to-incidence ratio (MIR) by age groups in high-income and low-income and middle-income countries, and globally in 1990, 2005, and 2010

<table>
<thead>
<tr>
<th>Age Group</th>
<th>1990</th>
<th>2005</th>
<th>2010</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Rate (95% CI)</td>
<td>n</td>
<td>Rate (95% CI)</td>
</tr>
<tr>
<td>&lt;75 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incidence</td>
<td>6 353 868</td>
<td>159·22 (145·32–174·98)</td>
<td>9 288 048</td>
<td>167·45 (150·96–187·11)</td>
</tr>
<tr>
<td>Prevalence</td>
<td>13 234 062</td>
<td>324·26 (288·74–374·96)</td>
<td>20 187 246</td>
<td>358·58 (317·58–412·79)</td>
</tr>
<tr>
<td>MIR</td>
<td>0·359 (0·318–0·409)</td>
<td>..</td>
<td>0·293 (0·249–0·332)</td>
<td>..</td>
</tr>
<tr>
<td>DALYs lost</td>
<td>63 991 864</td>
<td>1543·96 (1452·03–1728·25)</td>
<td>74 855 520</td>
<td>1326·17 (1172·08–1388·74)</td>
</tr>
<tr>
<td>Mortality</td>
<td>2 301 435</td>
<td>57·38 (54·12–64·27)</td>
<td>2 734 251</td>
<td>49·16 (43·60–51·55)</td>
</tr>
<tr>
<td>≥75 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incidence</td>
<td>3 725 067</td>
<td>3173·5 (2932·14–3422·23)</td>
<td>5 446 077</td>
<td>3082·97 (2819·52–3372·55)</td>
</tr>
<tr>
<td>Prevalence</td>
<td>4 681 276</td>
<td>3974·37 (3609·66–4441·23)</td>
<td>8 308 337</td>
<td>4700·18 (4239·37–5256·84)</td>
</tr>
<tr>
<td>MIR</td>
<td>0·634 (0·575–0·709)</td>
<td>..</td>
<td>0·543 (0·476–0·607)</td>
<td>..</td>
</tr>
<tr>
<td>DALYs lost</td>
<td>22 018 520</td>
<td>18665·35 (17 464·55–20 408·51)</td>
<td>27 096 178</td>
<td>15 300·36 (13 987·78–16 317·62)</td>
</tr>
<tr>
<td>Mortality</td>
<td>2 359 013</td>
<td>2033·21 (1888·78–2233·65)</td>
<td>2 950 719</td>
<td>1678·65 (1528·60–1807·22)</td>
</tr>
<tr>
<td>All ages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incidence</td>
<td>10 078 935</td>
<td>250·55 (229·70–273·25)</td>
<td>14 734 124</td>
<td>255·79 (232·10–283·88)</td>
</tr>
<tr>
<td>Prevalence</td>
<td>17 915 338</td>
<td>434·86 (389·45–496·84)</td>
<td>28 949 582</td>
<td>490·13 (436·60–557·52)</td>
</tr>
<tr>
<td>MIR</td>
<td>0·461 (0·415–0·518)</td>
<td>..</td>
<td>0·386 (0·336–0·432)</td>
<td>..</td>
</tr>
<tr>
<td>DALYs lost</td>
<td>86 010 384</td>
<td>2062·74 (1949·53–2280·29)</td>
<td>101 951 696</td>
<td>1749·59 (1568·67–1830·82)</td>
</tr>
<tr>
<td>Mortality</td>
<td>4 660 449</td>
<td>117·25 (111·51–129·68)</td>
<td>5 684 970</td>
<td>98·53 (89·02–103·86)</td>
</tr>
</tbody>
</table>

*p value for the difference in age-adjusted rates between 1990 and 2010 only.

Feigen et al., Lancet 383:245-255, 2014, Table 1
<table>
<thead>
<tr>
<th></th>
<th>1990</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td>(Continued from previous page)</td>
<td></td>
</tr>
<tr>
<td><strong>Globally</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;75 years</td>
<td></td>
</tr>
<tr>
<td>Incidence</td>
<td>6353868</td>
</tr>
<tr>
<td>Prevalence</td>
<td>13234062</td>
</tr>
<tr>
<td>MIR</td>
<td>..</td>
</tr>
<tr>
<td>DALYs lost</td>
<td>63991864</td>
</tr>
<tr>
<td>Mortality</td>
<td>2301435</td>
</tr>
</tbody>
</table>
What was wrong with that?

- **Way** too many digits.
- Numbers aren’t aligned.
- Numbers to be compared aren’t anywhere near each other.
- The interesting comparisons are horizontal rather than vertical.
- It would be much better as a multi-panel figure.
One last example

Drivers Involved In Fatal Collisions Who Were Not Distracted
As a share of the number of fatal collisions per billion miles, 2012

Drivers Involved In Fatal Collisions While Speeding
As a share of the number of fatal collisions per billion miles, 2009

Drivers Involved In Fatal Collisions While Alcohol-Impaired
As a share of the number of fatal collisions per billion miles, 2012

fivethirtyeight.com/datalab/which-state-has-the-worst-drivers
An alternative
Scatterplots

Non-distracted

Speeding

Alcohol

Ave Ins Premium

Ave Ins Loss

Premium vs Loss
• Show the data

• Avoid chart junk

• Consider taking logs and/or differences

• Put the things to be compared next to each other

• Use color to set things apart, but consider color blind folks

• Use position rather than angle or area to represent quantities
• Align things vertically to ease comparisons
• Use common axis limits to ease comparisons
• Use labels rather than legends
• Sort on meaningful variables (not alphabetically)
• Must 0 be included in the axis limits?
• Use scatterplots to explore relationships
Inspirations

• Hadley Wickham  *(slides at http://courses.had.co.nz)*
• Naomi Robbins  *(Creating more effective graphs)*
• Howard Wainer
• Andrew Gelman
• Dan Carr
• Edward Tufte
Further reading

The top ten worst graphs

With apologies to the authors, we provide the following list of the top ten worst graphs in the scientific literature.

As these examples indicate, good scientists can make mistakes.

bit.ly/TopTenWorstGraphs
Broman et al., Am J Hum Genet 63:861-869, 1998, Fig. 1
Kim et al., Nutr Res Pract 6:120-125, 2012 Fig 1

LDL phenotype B (n=13)
mean size: 248.2 Å (219.4~254.57)

LDL phenotype A (n=44)
mean size: 269.7 Å (257.34~294.63)
Distribution of All TFBS Regions

- Pseudogene/ambiguous: 17%
- 5' to known gene: 22%
- Novel: 24%
- Within or 3' flanking to a known gene: 36%

866 Total TFBS Regions

Cawley et al., Cell 116:499-509, 2004, Fig 1
Wittke-Thompson et al., Am J Hum Genet 76:967-986, Fig 1