Topics in Performance Evaluation

Dror Feitelson
Hebrew University

Lecture 2 – Graphs
“Few of us escape being indoctrinated with these notions:
(1) Numerical calculations are exact, but graphs are rough;
(2) For any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
(3) Performing intricate calculations is virtuous, whereas actually looking at the data is cheating.”

F. J. Anscombe

*The American Statistician* **27**(1) Feb 1973
Anscombe's example of 4 datasets:

<table>
<thead>
<tr>
<th>x1</th>
<th>y1</th>
<th>x2</th>
<th>y2</th>
<th>x3</th>
<th>y3</th>
<th>x4</th>
<th>y4</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
<td>13</td>
<td>12.74</td>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>5.39</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
<td>12.5</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>8.15</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>6.42</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>5.73</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
<td>6.89</td>
</tr>
</tbody>
</table>

What can you say about them?
Let's calculate some descriptive statistics for dataset #1:

- number of observations: 11
- mean of $x$: 9.0
- mean of $y$: 7.5
- linear regression: $y = 3 + 0.5x$
- $R^2$: 0.667
- correlation coefficient: 0.82
- sum of squares of $x-\text{avg}(x)$: 110.0
- regression sum of squares: 27.5
- residual sum of squares of $y$: 13.75
- estimated std. error of slope: 0.118

For the other data sets we get the same results!!!
so they are all similar, right?
Conclusion:

look at the data!

- Discover what the data has to say
  John W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977
- Display your conclusions in the most convincing manner
Graphs that made history or illuminate data

Michael Friendly's Gallery of data visualization
William Playfair, *The Commercial and Political Atlas*, 1786: invented most graphs used today
Charles Minard, plot of Napoleon's failed campaign in Russia, 1812

Popularized by Tufte as the best graphic ever
John Snow, deaths in London Cholera epidemic, 1854

Established link between water quality and health

Precursor of modern GIS systems
Florence Nightingale, British casualties in Crimean war, 1858

Established sanitation as a decisive factor in hospital operation

From: F. Nightingale, "Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army", 1858
Summary of a whole year's weather

- Lots of numbers (daily max/min + average max/min + humidity)
- Use of parallel graphs for correlation
- Callouts to emphasize special points
World Health Chart 2001

Download from www.whc.ki.se the beta version of a free software showing World Health in motion towards Millennium Development Goals.

Data from all 174 countries & territories with > 250,000 inhabitants
Source: World Development Indicators 2002 and estimates in italic
© Hans Rosling, hans.roslings@phs.ki.se
Division of International Health, Dept. of Public Health Sciences,
Karolinska Institutet, SE-171 76, Stockholm, Sweden
Road casualty data relative to population.
Radial sections are pedestrians, cars, and motorcycle
"Academic" Salaries

Actual average and median salaries at U.S. Doctoral-granting Universities

Notes: Administrator figures are medians salaries, the rest are averages. All figures in 2008 dollars. Sources: College and University Professional Association for Human Resources 2005 Survey; American Association of University Professors 2007 Survey; The Chronicle of Higher Education 2001 Survey of Graduate Assistants; USA Today Survey of Div. I-A College Football Coaches Compensation 2007.
The harm of bad graphics

Tufte, *Visual Explanations* using graphs by Morton Thiokol, Inc.
Background: launch of the Challenger space shuttle on 27 January 1986, amid concerns regarding O-ring function in cold weather.
Data regarding test rockets from the manufacturer (chart prepared later; charts used in discussions prior to the launch contained less data)

**History of O-Ring Damage in Field Joints**

<table>
<thead>
<tr>
<th>Development Motor Number</th>
<th>O-Ring Temp (°F)</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64°</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>49°</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>61°</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>40°</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>58°</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Qualification Motor Number</th>
<th>O-Ring Temp (°F)</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89°</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>67°</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>45°</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>60°</td>
<td></td>
</tr>
</tbody>
</table>

**Code**
- **S**: Heating of Secondary O-Ring
- **B**: Primary O-Ring Blowby
- **E**: Primary O-Ring Erosion
- **H**: Heating of Primary O-Ring
- **X**: No Damage

**STATIC TEST MOTORS**
- **HORIZONTAL ASSEMBLY**
- **SOME PUTTY REPAIRED**
Data regarding prior launches
Note that legend is missing (appeared previously)

Contains too much irrelevant data

Does not clarify effect of temperature
Tufte's alternative rendering of the data

O-ring damage index, each launch

26°-29° range of forecasted temperatures (as of January 27, 1986) for the launch of space shuttle Challenger on January 28

Note exaggerated X scale for emphasis
Examples
A graph should be independent and provide full information

- Title (if relevant)
- Axis labels (including units)
- Tics indicating values
- Legend

![](image-url)
Need to also consider aesthetics

- **Proportions**
  - Size and placement of labels and legend
  - Size of fonts relative to graphical elements

- **Use of color**
  - Express gradient with deeper shades
  - Create focus for discussion
  - Should also work in black and white

- **Combination of graphical elements**
  - Give full picture
  - Connections through consistent use of colors

- **Order in legend matches order of graphs**
Causal/functional relationship:

- XY plot (continuous)
- bar chart (categories)
- scatter plot (complicated)
Showing measurements

- Emphasize points
- Connect with weaker lines
Or show fitted model line here model is \( y \sim 1/x \) often linear regression
Scales

- Linear is best
- Logarithmic if needed
Scales

- Logarithmic if needed
- Show values, not their log, in stubs
Log scale

- Beware of expansion in small values
Stubs

- Uniform scale (Y axis)
- Match measured values (powers of 2)
- Show important values (maximal size)
Axis break
useful for few extreme values
Stacking
Show individual components and also their sum
Histogram: number of samples between $x$ and $x+\delta$

- Simplest display of a distribution
- Sensitive to bin size $\delta$
CDF: Probability of sample smaller than $x$

- Robust
- Alternative to histogram

![Graphs showing the CDF for different time intervals: 10 seconds, 1 minute, 10 minutes, 30 minutes, 1 hour, and 2.5 hours. Each graph represents the cumulative distribution of thousand jobs up to a given hour.]
CDF
- Modes less prominent
- Easier to see where weight is concentrated
Box plot

- Summary of a distribution
Comparison of distributions

Graphs showing the distribution of interarrival times for different users and groups on SDSC SP2.
Comparison of distribution of results for different experimental parameter values
Skewed distributions are common
note difference between mean and median