JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

How to display data badly

Karl W Broman
Department of Biostatistics

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

Protecting Health, Saving Lives—*Millions at a Time*

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Using Microsoft Excel to obscure your data and annoy your readers

Karl W Broman
Department of Biostatistics

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

Protecting Health, Saving Lives—*Millions at a Time*

Inspiration

This lecture was inspired by

H Wainer (1984) How to display data badly. 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.

3

General principles

The aim of good data graphics:

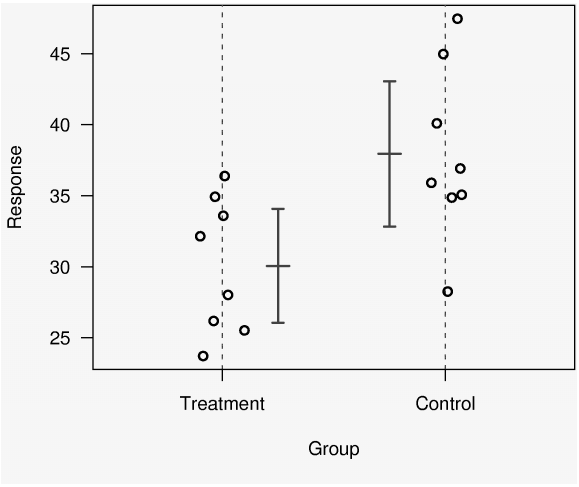
Display data accurately and clearly.

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.

4

Example 1



5

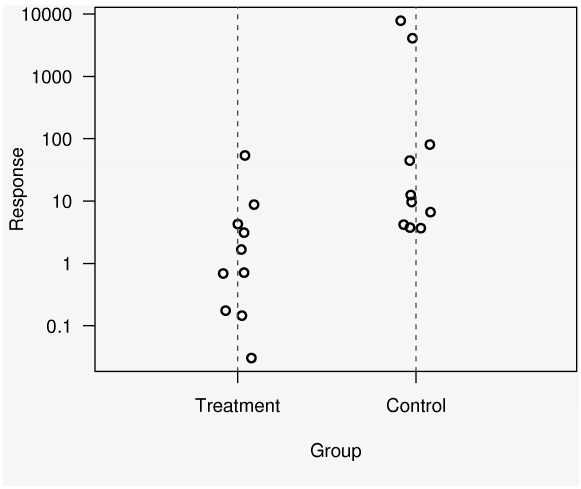
Example 2

Distribution of genotypes

AA	21%
AB	48%
BB	22%
missing	9%

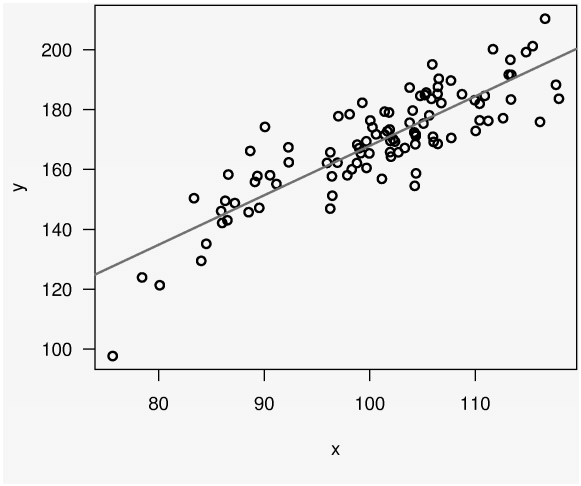
6

Example 3



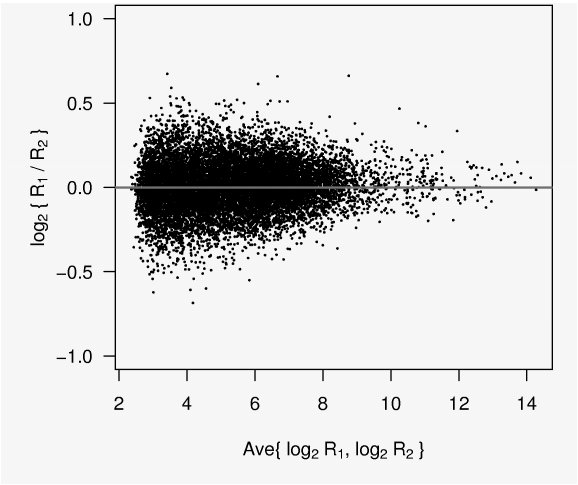
7

Example 4



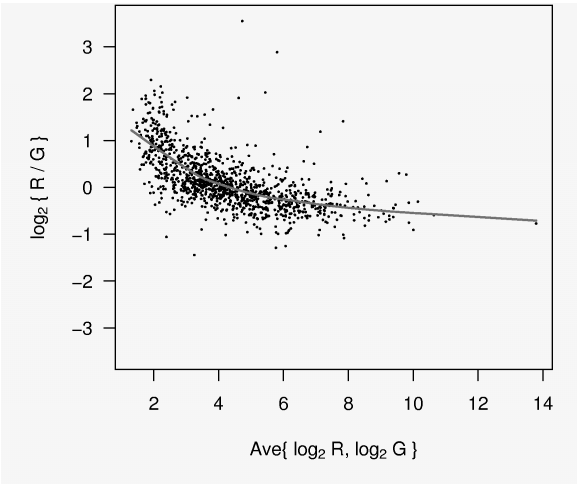
8

Example 5



9

Example 6



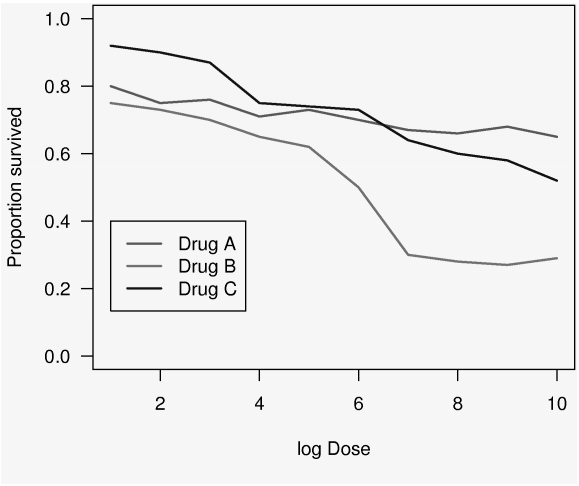
10

Example 7

	$b/c = 10.0$		$b/c = 10.0$		$b/c = 100.0$	
N	$r^\star$	G	$r^\star$	G	$r^\star$	G
3	2	0.20	2	2.2	2	22
4	2	0.26	2	2.9	2	29
5	2	0.32	3	3.5	3	36
6	3	0.38	3	4.2	3	43
7	3	0.45	3	4.9	3	49
8	3	0.51	4	5.6	4	56
9	3	0.57	4	6.3	4	63
10	4	0.63	4	6.9	4	70

11

Example 8



12

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.

13

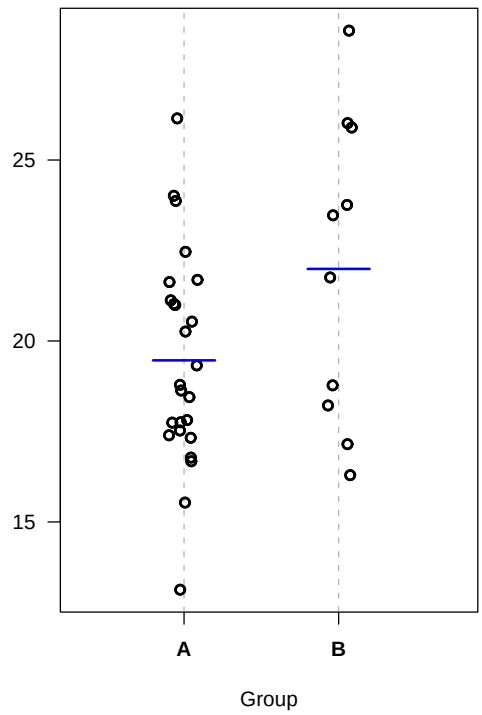
Further reading

- ER Tufte (1983) The visual display of quantitative information. Graphics Press.
- ER Tufte (1990) Envisioning information. Graphics Press.
- ER Tufte (1997) Visual explanations. Graphics Press.
- WS Cleveland (1993) Visualizing data. Hobart Press.
- WS Cleveland (1994) The elements of graphing data. CRC Press.

14

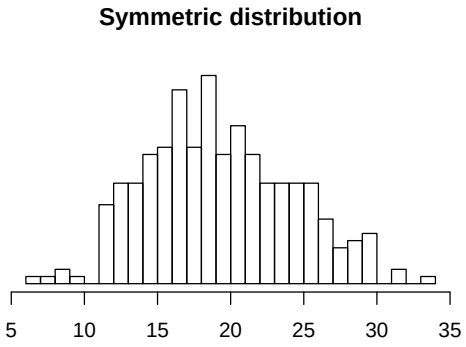
Displaying distributions

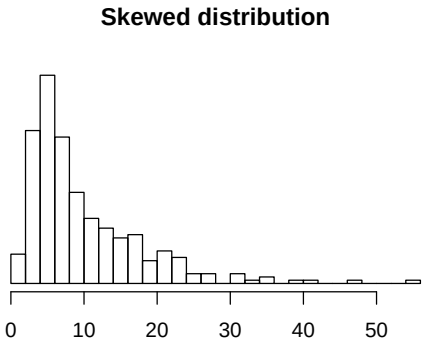
Dotplots



- Few data points per group
- Possibly many groups

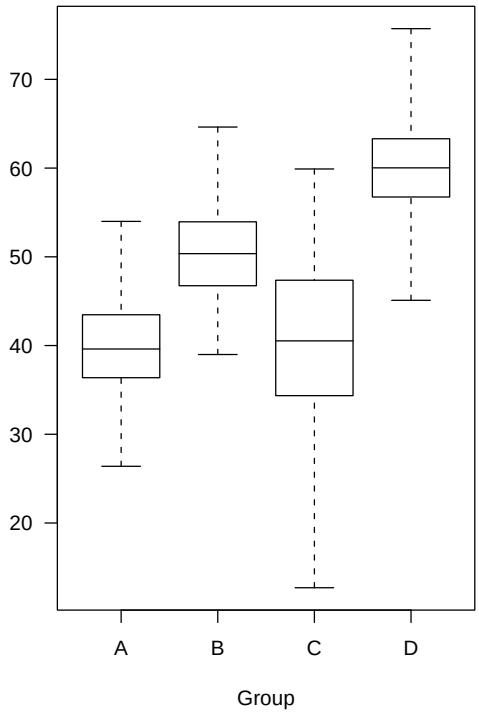
Histograms





- 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

Location/ Center	mean (i.e., average) median mode geometric mean harmonic mean
Scale	standard deviation (SD) inter-quartile range (IQR) range
Other	quantile quartile quintile

The means

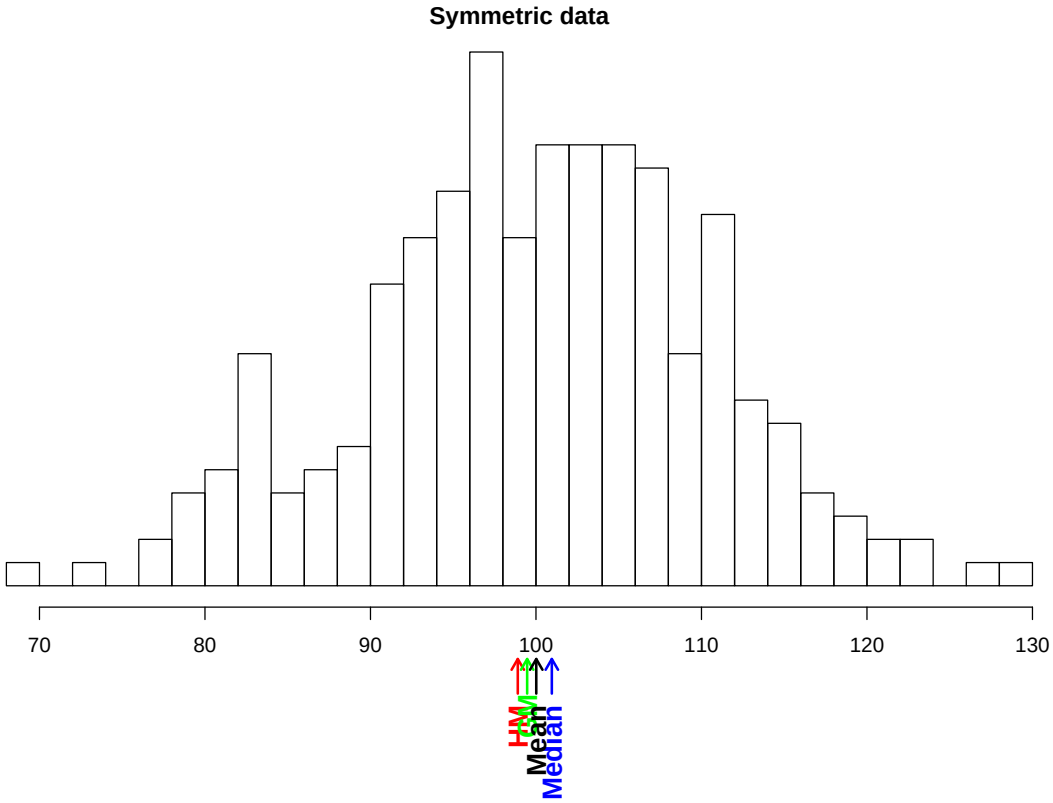
$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = (x_1 + x_2 + \ldots + x_n)/n$$

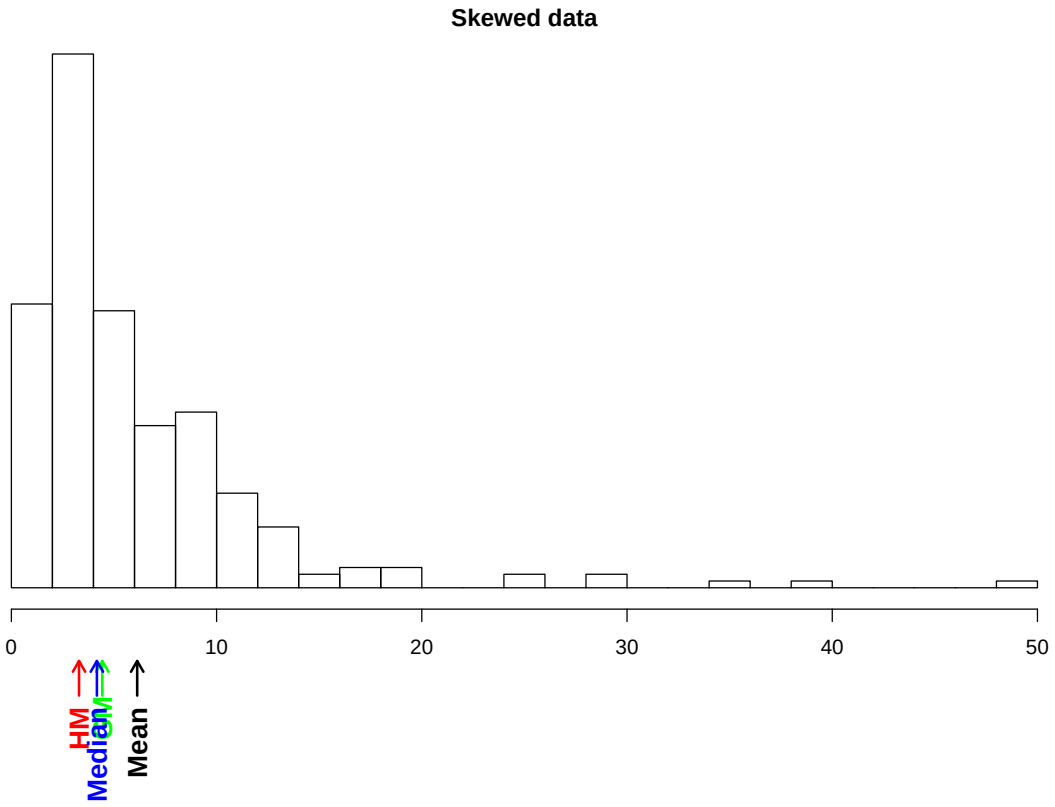
$$\text{geometric mean} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} \log x_i\right\}$$

$$\text{harmonic mean} = 1/\left\{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)\right\}$$

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 $= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2$

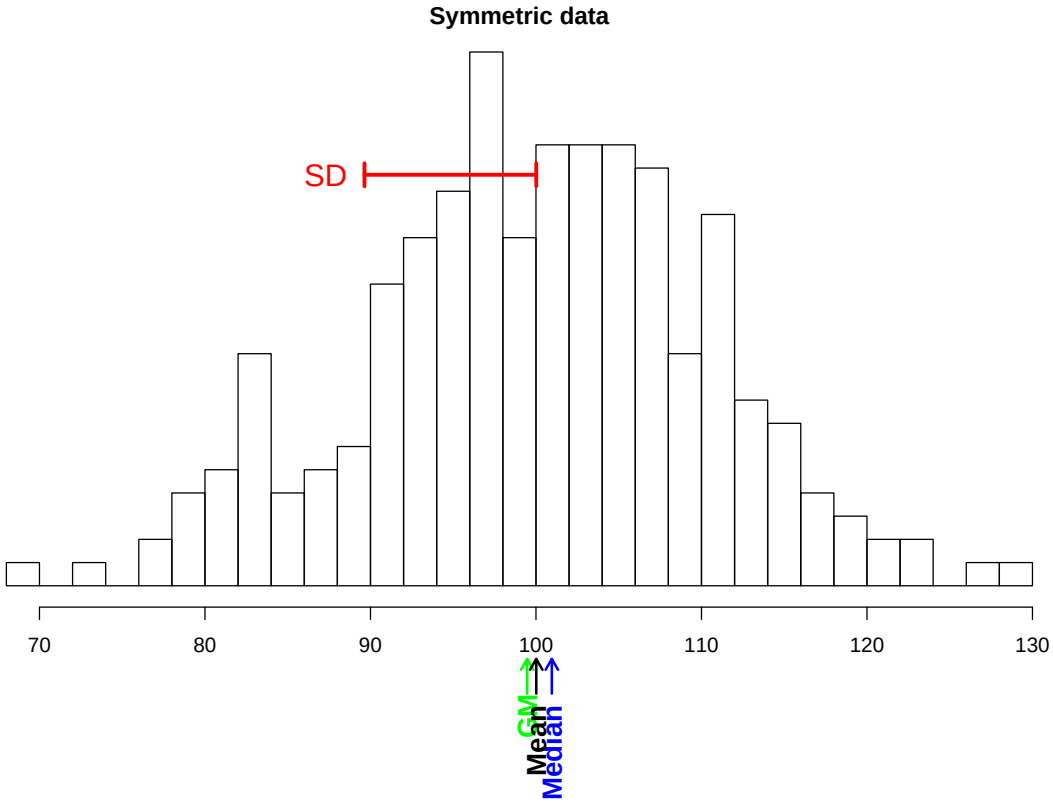
sample SD = $\sqrt{s^2} = s$

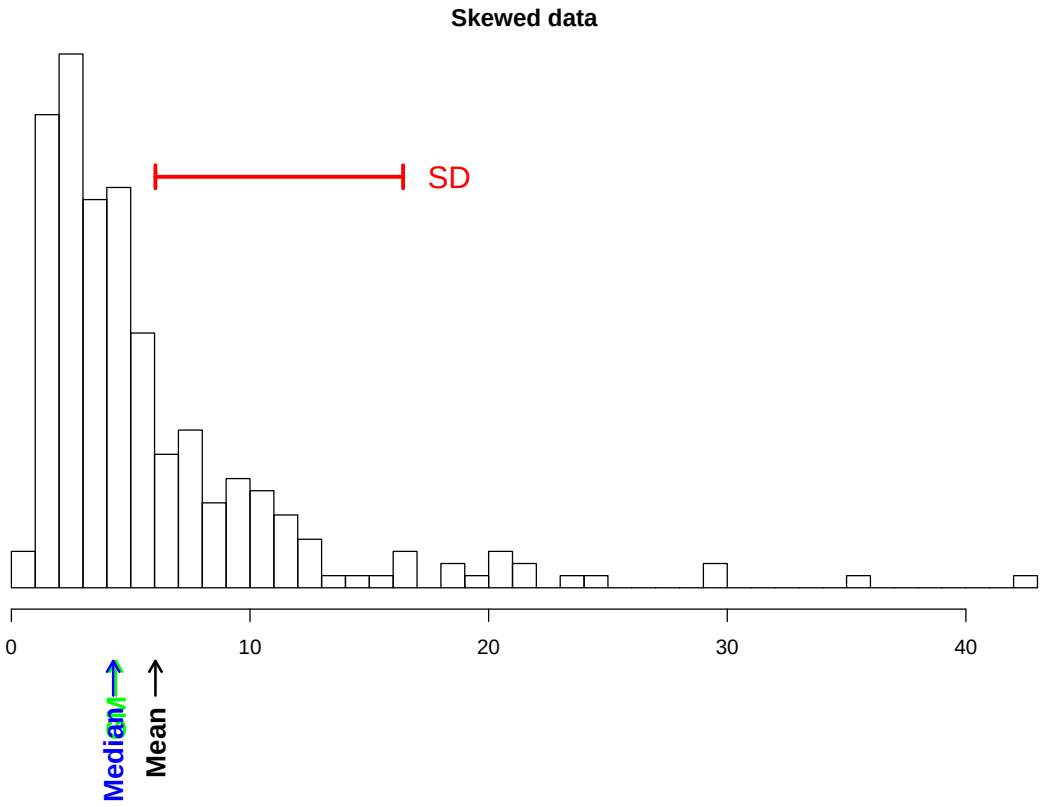
= RMS (distance from average)

= “typical” distance from the average

= sort of like ave$\{|x_i - \bar{x}|\}$

Note: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$




Skewed data
SD
0
10
20
30
40
Median
Mean