Visualisation of quantitative information

James Neill, 2011

Overview

1. Visualisation
2. Approaching data
3. Levels of measurement
4. Principals of graphing
5. Univariate graphs
6. Graphical integrity

Visualisation

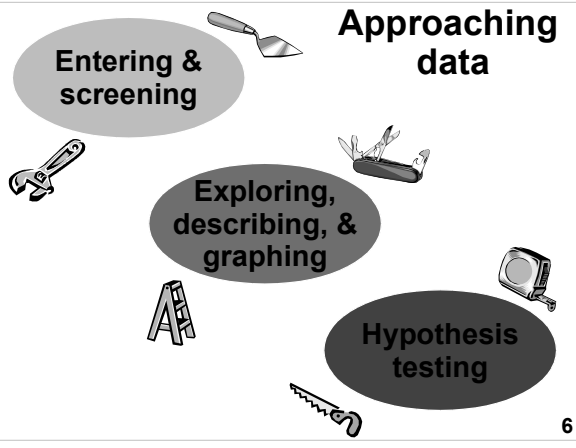
"Visualization is any technique for creating images, diagrams, or animations to communicate a message." - Wikipedia

Is Pivot a turning point for web exploration?

(Gary Flake)

(TED talk - 6 min.)

Approaching data



Describing & graphing data

THE CHALLENGE:
to find a meaningful, accurate
way to depict the
'true story' of the data

7

Get your fingers dirty with data

8

Get intimate with your data

9

Clearly report the data's main features

10

Levels of Measurement = Type of Data

Stevens (1946)

11

Levels of measurement

- **N**ominal / Categorical
- **O**rdinal
- **I**nterval
- **R**atio

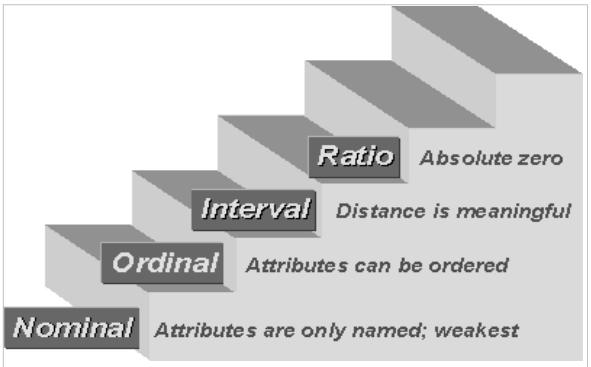
12

Discrete vs. continuous

Discrete

- - - - - - - - - -

Continuous



Each level has the properties of the preceding levels, plus something more!

Categorical / nominal

- Conveys a category label
- (Arbitrary) assignment of #s to categories
  e.g. Gender

Male	Female
0	1
🚹	🚺

- *No useful information,* except as labels

Ordinal / ranked scale

- Conveys *order*, but not *distance*
  e.g. in a race, 1st, 2nd, 3rd, etc. or ranking of favourites or preferences

Ordinal / ranked example:
Ranked importance

Rank the following aspects of the university according to what is most important to you (1 = most important through to 5 = least important)

__ Quality of the teaching and education
__ Quality of the social life
__ Quality of the campus
__ Quality of the administration
__ Quality of the university's reputation

Interval scale

- Conveys *order* & *distance*
- 0 is arbitrary
  e.g., temperature (degrees C)
- Usually treat as continuous for > 5 intervals

Interval example:
8 point Likert scale

FALSE NOT LIKE ME — TRUE LIKE ME

1	2	3	4	5	6	7	8
This statement doesn't describe me at all; it isn't like me at all		More false than true		More true than false		This statement describes me very well; it is very much like me.	

SOME EXAMPLES

A. I am a fast thinker. 1 2 3 4 5 6 (7) 8
(The 6 has been circled because the person answering believes the statement "I am a fast thinker" is sometimes true. That is, the statement is sometimes like him/her.)

B. I am a good storyteller. 1 (2) 3 4 5 6 7 8
(The 2 has been circled because the person answering believes that the statement is mostly false as far as he/she is concerned. That is, he/she feels he/she does not tell good stories.)

C I enjoy working on puzzles. 1 2 3 4 5 6 7 (8)
(The 8 has been circled because the person really enjoys working on puzzles a great deal, therefore the statement is definitely true about him/her.)

Ratio scale

- Conveys *order* & *distance*
- Continuous, with a meaningful 0 point
  e.g. height, age, weight, time, number of times an event has occurred
- Ratio statements can be made
  e.g. X is twice as old (or high or heavy) as Y

20

Ratio scale:
Time

Estimate **the average hours per week** (approx.) you spend during semester:

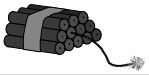
10. in paid employment _____
11. in classes (lectures, tutorials etc.) _____
12. studying outside of classes _____

21

Why do levels of measurement matter?

Different analytical procedures are used for different levels of data.

More powerful statistics can be applied to higher levels



22

Principles of graphing

23

Graphs
(Edward Tufte)

- Visualise data
- Reveal data
  - Describe
  - Explore
  - Tabulate
  - Decorate
- Communicate complex ideas with clarity, precision, and efficiency

24

Tufte's graphing guidelines

- Show the data
- Avoid distortion
- Focus on substance rather than method
- Present many numbers in a small space
- Make large data sets coherent

Tufte's graphing guidelines

- Maximise the information-to-ink ratio
- Encourage the eye to make comparisons
- Reveal data at several levels/layers
- Closely integrate with statistical and verbal descriptions

Graphing steps

1. Identify the purpose of the graph
2. Select which type of graph to use
3. Draw a graph
4. Modify the graph to be clear, non-distorting, and well-labelled.
5. Disseminate the graph (e.g., include it in a report)

Software for data visualisation (graphing)

1. Statistical packages
- e.g., SPSS

2. Spreadsheet packages
- e.g., MS Excel

3. Word-processors
- e.g., MS Word – Insert – Object – Micrograph Graph Chart

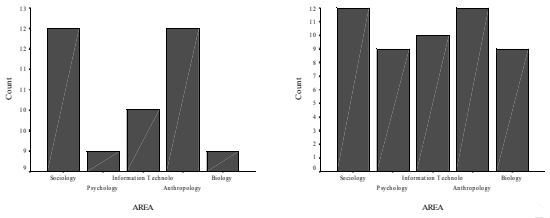
Univariate graphs

Univariate graphs

- Bar graph
- Pie chart
- Data plot
- Error bar
- Stem & leaf plot
- Box plot (Box & whisker)
- Histogram

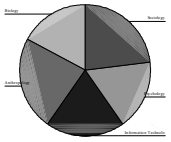
Bar chart (Bar graph)

- Examine comparative heights of bars
- X-axis: Collapse if too many categories
- Y-axis: Count or % or mean?
- Consider whether to use data labels



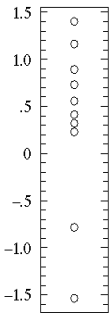
Pie chart

- Use a bar chart instead
- Hard to read
  - –Does not show small differences
  - –Rotation / position influences perception

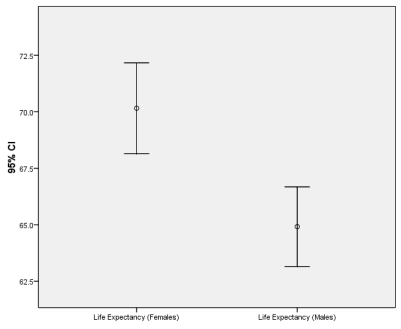


Data plot & error bar

Data plot



Error bar



Stem & leaf plot

- Alternative to histogram
- Use for ordinal, interval and ratio data
- May look confusing to unfamiliar reader

Raw Data	Stem	Leaf
0 1 1 2 2 3 4 4 4 5 5 5 6 6 7 7 7 7 8 8 9 9	0	01122344455566777788899
	1	011122233333444555556666666666777888899
10 11 11 11 12 12 12 13 13 13 13 14 14 14 15 15 15 15 15 16 16 16 16 16 16 16 16 16 17 17 17 18 18 18 18 19 19	2	00112233444455667889
	3	005
20 20 21 21 22 22 23 23 24 24 24 24 25 25 26 26 27 28 28 29		
30 30 35		

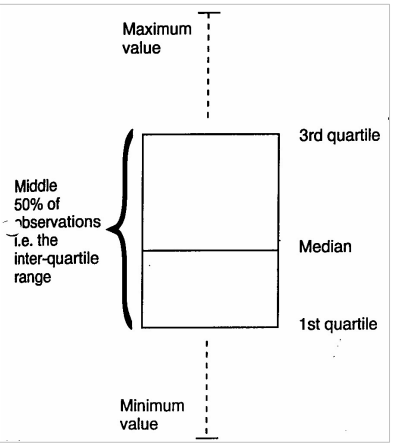
Stem & leaf plot

- Contains actual data
- Collapses tails

```
Frequency     Stem &  Leaf
     7.00      1 .  &
   192.00      1 .  22223333333
   541.00      1 .  44444444444444445555555555555
   610.00      1 .  66666666666666777777777777777777
   849.00      1 .  88888888888888888888888888999999999999999999
   614.00      2 .  00000000000000001111111111111111
   602.00      2 .  22222222222222233333333333333333
   447.00      2 .  44444444444445555555555
   291.00      2 .  66666666677777777
   240.00      2 .  88888889999999
   167.00      3 .  000001111
   146.00      3 .  22223333
   153.00      3 .  44445555
   118.00      3 .  666777
    99.00      3 .  888999
   106.00      4 .  000111
    54.00      4 .  222
   339.00 Extremes    (>=43)
```

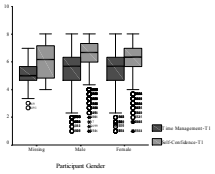

Box plot (Box & whisker)

- Useful for interval and ratio data
- Represents min., max, median, quartiles, & outliers



Box plot

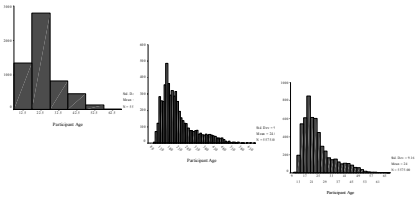
- Alternative to histogram
- Useful for screening
- Useful for comparing variables
- Can get messy - too much info
- Confusing to unfamiliar reader



Participant Gender

Histogram

- For continuous data
- X-axis needs a happy medium for # of categories
- Y-axis matters (can exaggerate)



Histogram of male & female heights

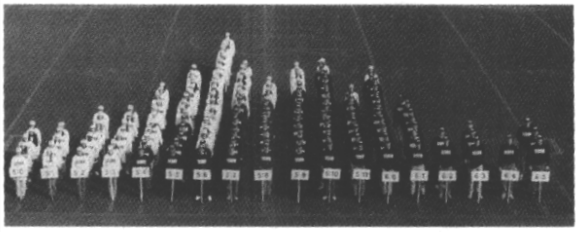


FIGURE 2.3.11 Histogram of heights constructed using the people. Photograph by Peter Morenus in conjunction with Prof. Linda Strausberg, University of Connecticut. Subjects are University of Connecticut genetics students, females in white tops, males in dark tops.

Non-normal distributions

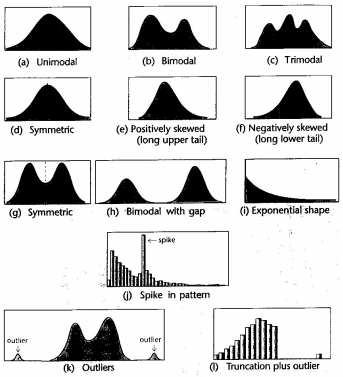
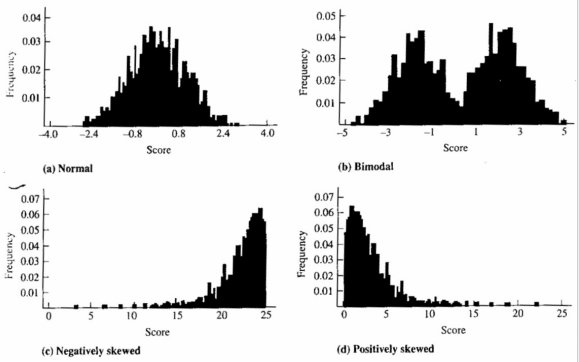
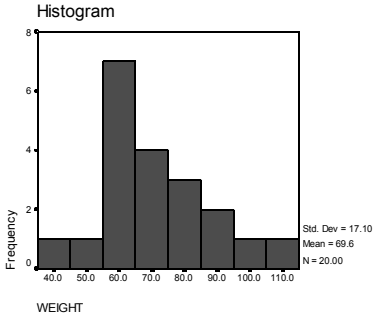


FIGURE 2.3.10 Features to look for in histograms and stem-and-leaf plots.

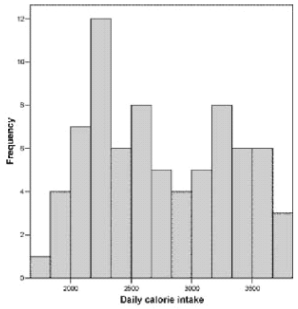
Non-normal distributions



Histogram of weight

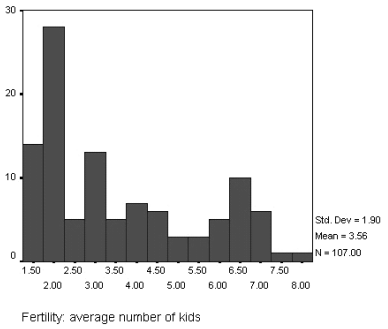


Histogram of daily calorie intake

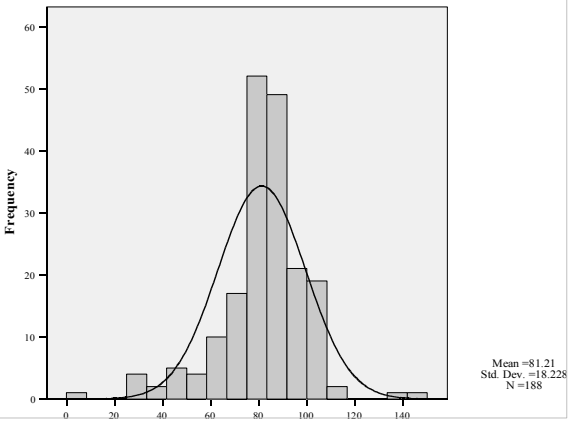


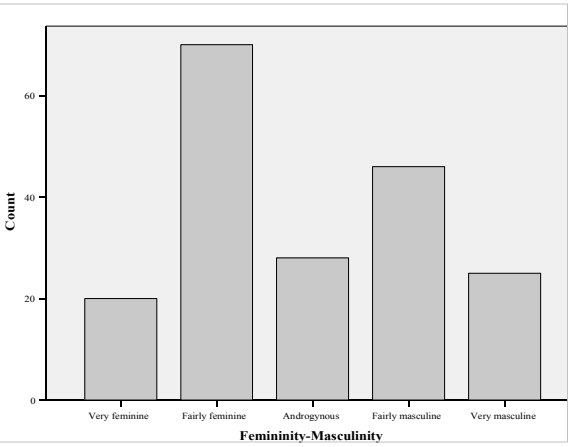
43

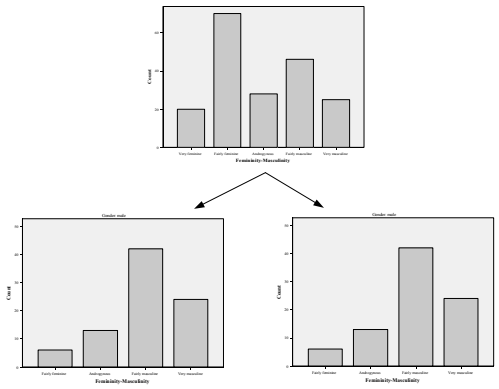
Histogram of fertility



44

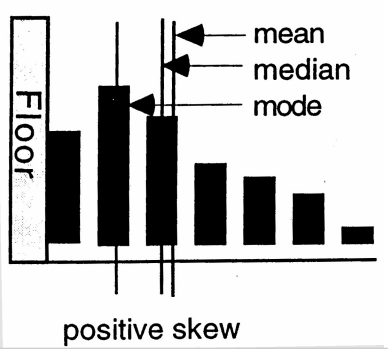






47

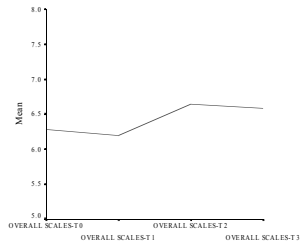
Effects of skew on measures of central tendency



48

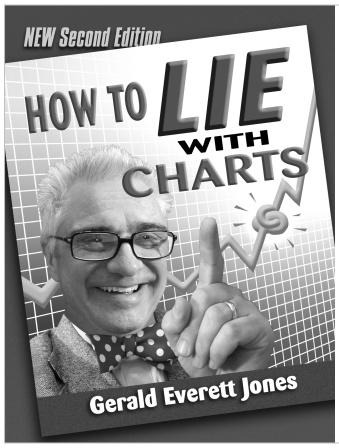
Line graph

- Alternative to histogram
- Implies continuity e.g., time
- Can show multiple lines



Summary: Graphs & levels of measurement

	NOIR
Bar chart & pie chart	NOI
Histogram	IR
Stem & leaf	IR
Data plot & box plot	IR
Error-bar	IR
Line graph	IR



Graphical integrity

(part of academic integrity)

Graphing can be like a bikini. What they reveal is suggestive, but what they conceal is vital.

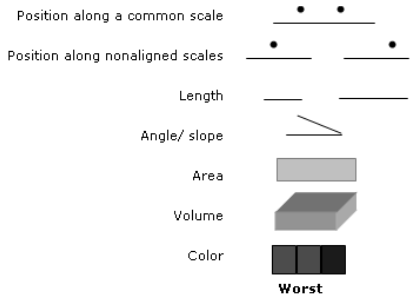
(aka Aaron Levenstein)

"Like good writing, good graphical displays of data communicate ideas with clarity, precision, and efficiency.

Like poor writing, bad graphical displays distort or obscure the data, make it harder to understand or compare, or otherwise thwart the communicative effect which the graph should convey."

Michael Friendly –
Gallery of Data Visualisation

Cleveland's hierarchy



Based on graphic (Figure 2) in Presentation Graphics (white paper) by Leland Wilkinson, SPSS, Inc and Northwestern Uiv.

Cleveland's hierarchy:
Best to worst

1. Position along a common scale
2. Position along identical, non aligned scales
3. Length
4. Angle-slope
5. Area
6. Volume
7. Color hue - color saturation - density

Tufte's graphical integrity

- Some lapses intentional, some not
- Lie Factor = size of effect in graph / size of effect in data
- Misleading uses of area
- Misleading uses of perspective
- Leaving out important context
- Lack of taste and aesthetics

Review questions

1. If a survey question produces a 'floor effect', where will the mean, median and mode lie in relation to one another?
2. Over the last century, the performance of the best baseball hitters has declined. Does this imply that the overall performance of baseball batters has decreased?

Can you complete this table?

Level	Properties	Examples	Descriptive Statistics	Graphs
Nominal /Categorical				
Ordinal / Rank				
Interval				
Ratio				

Answers: http://wilderdom.com/research/Summary_Levels_Measurement.html

Links

- Presenting Data – Statistics Glossary v1.1 - http://www.cas.lancs.ac.uk/glossary_v1.1/presdata.html
- A Periodic Table of Visualisation Methods - http://www.visual-literacy.org/periodic_table/periodic_table.html
- Gallery of Data Visualization - http://www.math.yorku.ca/SCS/Gallery/
- Univariate Data Analysis – The Best & Worst of Statistical Graphs - http://www.csulb.edu/~msaintg/ppa696/696uni.htm
- Pitfalls of Data Analysis – http://www.vims.edu/~david/pitfalls/pitfalls.htm
- Statistics for the Life Sciences – http://www.math.sfu.ca/~cschwarz/Stat-301/Handouts/Handouts.

References

1. Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.
2. Jones, G. E. (2006). *How to lie with charts*. Santa Monica, CA: LaPuerta.
3. Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Open Office Impress

- This presentation was made using Open Office Impress.
- Free and open source software.
- http://www.openoffice.org/product/impress.html