Descriptives & Graphing

Lecture 3

Survey Research & Design in Psychology

James Neill, 2012

Overview: Descriptives & Graphing

1. Approaching data
2. Descriptive statistics
3. Normal distribution
4. Non-normal distributions
5. The effect of skew on central tendency
6. Graphical techniques

Is Pivot a turning point for web exploration?

(Gary Flake)

(TED talk - 6 min.)

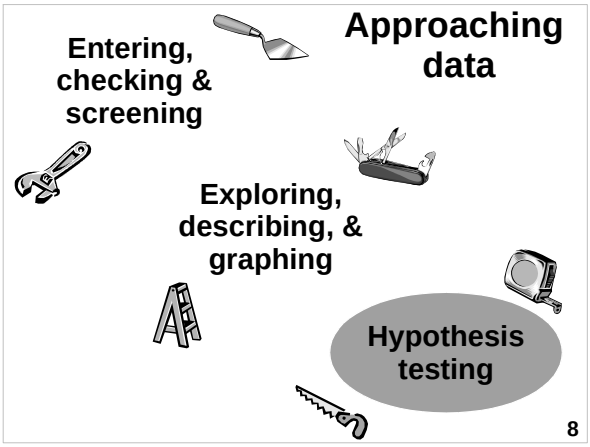
Approaching data

Get your fingers dirty with data

Get intimate with your data

Clearly report the data's main features

7



8

Data checking

- Hard copy survey data can be checked by having one person read the survey responses aloud to another person checking the data file.
- A percentage of the surveys can be checked for a large study.
- Report error-rate in research report

9

Data screening

- Carefully 'screen' a data file to help minimise errors and maximise validity.
- Out of range and mis-entered data
- Missing and duplicate cases
- Missing data
- Recoding

10

Exploring, describing & graphing data

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

11

Parametric & non-parametric statistics & level of measurement

12

Level of measurement determines type of descriptive statistics and graphs

GOLDEN RULE of DATA ANALYSIS

The level of measurement (see previous lecture) determines which types of descriptive statistics and which types of graphs are appropriate.

Levels of measurement and non-parametric vs. parametric

Categorical & ordinal DVs
→ ***non-parametric***
(Does not assume a normal distribution)

Interval & ratio DVs
→ ***parametric***
(Assumes a normal distribution)
→ ***non-parametric***
(If distribution is non-normal)

DVs = dependent variables

Parametric statistics

- Procedures which estimate **parameters** of a population, usually based on the **normal distribution**
  - *M*, *SD*, skewness, kurtosis
    - *t*-tests, ANOVAs
  - *r*
    - bivariate correlation, linear regression

Parametric statistics

- More powerful (more sensitive)
- But they require more assumptions and are more vulnerable to violations of assumptions comapred to non-parametrics statistics

Non-parametric statistics

(Distribution-free tests)

- Procedures which do not rely on estimates of population parameters
  - Frequency
    - e.g. sign test, chi-squared
  - Rank order
    - e.g. Mann-Whitney U test, Wilcoxon matched-pairs signed-ranks test

Univariate descriptive statistics

Number of variables

Univariate
= one variable

e.g., mean, median, mode, histogram, bar chart, box plot

Bivariate
= two variables

e.g., correlation, *t*-test, scatterplot, clustered bar chart

Multivariate
= more than two variables

e.g., reliability analysis, factor analysis, multiple linear regression

What do we want *described*?

The **distributional properties** of underlying variables, based on:

- **Central tendency**(ies): Frequencies, Mode, Median, Mean
- **Shape**: Skewness, Kurtosis
- **Spread** (dispersion): Min., Max., Range, IQR, Percentiles, Var/SD

for sampled data.

Measures of central tendency

Statistics which represent the 'centre' of a frequency distribution:

- Mode (most frequent)
- Median (50th percentile)
- Mean (average)

Which ones to use depends on:

- Type of data (level of measurement)
- Shape of distribution (esp. skewness)

Reporting more than one may be appropriate.

Measures of central tendency

	Mode	Median	Mean
Nominal	√		
Ordinal	√	√	
Interval	√	√	√
Ratio	?	√	√

Measures of shape / spread / dispersion / deviation

- Measures of shape and deviation from the central tendency

Non-parametric / non-normal:

- Min and max
- Range
- Percentiles

Parametric:

- *SD*
- Skewness
- Kurtosis

Measures of spread / dispersion / deviation

	Min/Max, Range	Percentiles	*Var/SD*
Nominal			
Ordinal	√	√	
Interval	√	√	√
Ratio	√	√	√

Describing nominal data

- **Nominal** = Labelled categories
- Descriptive statistics:
  - –Most frequent? (Mode – e.g., females)
  - –Least frequent? (e.g, Males)
  - –Frequencies (e.g., 20 females, 10 males)
  - –Percentages (e.g. 67% females, 33% males)
  - –Cumulative percentages
  - –Ratios (e.g., twice as many females as males)

Describing ordinal data

- **Ordinal** = Conveys order but not distance (e.g., ranks)
- Descriptives approach is as for nominal (frequencies, mode etc.)
- Plus percentiles (including median) may be useful

Describing interval data

- **Interval** = order and distance, but no true 0 (0 is arbitrary).
- Central tendency (mode, median, mean)
- Shape/Spread (min, max, range, *SD*, skewness, kurtosis)

Interval data is discrete, but is often treated as ratio/continuous (especially for > 5 intervals)

Describing ratio data

- **Ratio** = Numbers convey order and distance, meaningful 0 point
- Descriptives approach is as for interval (i.e., median, mean, SD, skewness etc.)
- Ratios

Mode (*Mo*)

- Most common score - highest point in a frequency distribution – a real score – for most no. of participants
- Suitable for all levels of data, but may not be appropriate for ratio
- Not affected by outliers
- Check frequencies and bar graph to see whether it is an accurate and useful statistic.

Frequencies

- # of units in each category
- % of units in each category
- Frequency table
- Bar chart or pie graph
- Crosstabs (contingency table) is the bivariate equivalent of frequencies

Median (*Mdn*)

- Mid-point of distribution (Q2, 50th percentile)
- Not badly affected by outliers
- May not represent the central tendency in skewed data
- If the Median is useful, then consider what other percentiles may also be worth reporting.

Summary: Descriptive statistics principles

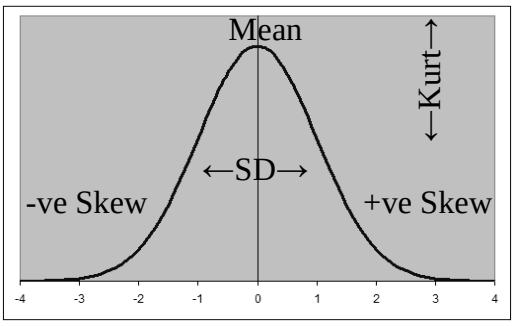
- Spend '**quality time**' investigating (exploring and describing) your data
- Describe the **central tendency**
  - Frequencies, Percentages
  - Mode, Median, Mean
- Describe the **variability**:
  - Min, Max, Range, Quartiles
  - Standard Deviation, Variance

Summary: Descriptive statistics & levels of measurement

	NOIR
Frequencies	NOI?
Mode	NOI?
Median	OIR
Mean	IR
Min-Max-IQR	OIR
SD	IR

Properties of the normal distribution

The four moments of a normal distribution



The four moments of a normal distribution

Four mathematical qualities (parameters) can describe a continuous distribution which as least roughly follows a bell curve shape:

- 1st = mean (central tendency)
- 2nd = *SD* (dispersion)
- 3rd = skewness (lean / tail)
- 4th = kurtosis (peakedness / flattness)

Mean (1st moment)

- Average score

$$\text{Mean} = \Sigma X / N$$

- Use for normally distributed ratio data or interval (if treating it as continuous).
- Influenced by extreme scores (outliers)

37

Beware inappropriate averaging...

With your head in an oven
and your feet in ice
you would feel,
on average,
just fine

The majority of people have more than the average number of legs (M = 1.9999).

38

Standard deviation (2nd moment)

- *SD* = square root of the variance

$$= \frac{\Sigma (X - X)^2}{N - 1}$$

- Used for normally distributed interval or ratio data
- Affected by outliers
- Standard Error (SE) = *SD* / square root of *N*

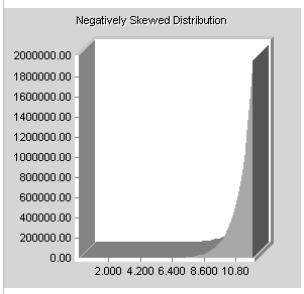
39

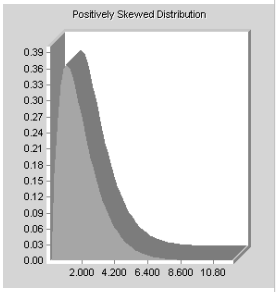
Skewness (3rd moment)

- Lean of distribution
  - +ve = tail to right
  - -ve = tail to left
- Can be caused by:
  - an outlier, or ceiling or floor effects
- Can be accurate
  - e.g., no. of cars owned per person

40

Skewness (3rd moment)
(with ceiling and floor effects)





- Negative skew
- Positive skew

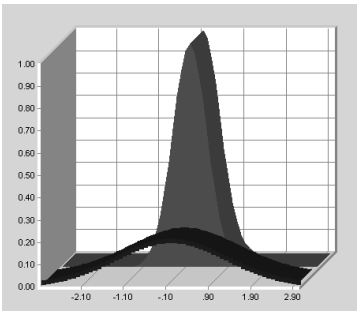
41

Kurtosis (4th moment)

- Flatness or peakedness of distribution
  - +ve = peaked
  - -ve = flattened
- By altering the X &/or Y axis, any distribution can be made to look more peaked or flat – add a normal curve to help judge kurtosis visually.

42

Kurtosis (4th moment)



Red = Positive (leptokurtic)

Blue = Negative (platykurtic)

Judging severity of skewness & kurtosis

- View histogram with normal curve
- Deal with outliers
- Rule of thumb: Skewness and kurtosis > -1 or < 1 is generally considered to reasonable for parametric inferential statistics
- Significance tests: Tend to be overly sensitive

Areas under the normal curve

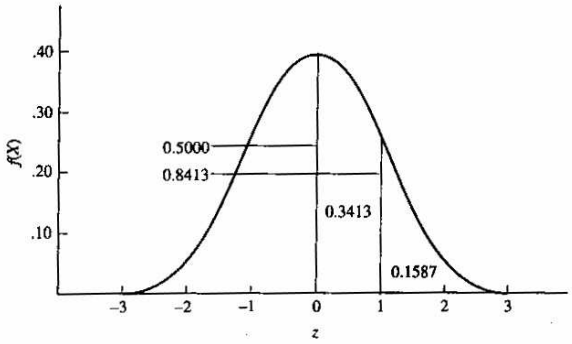
If distribution is normal (bell-shaped - or close):

~68% of scores within +/- 1 *SD* of *M*

~95% of scores within +/- 2 *SD* of *M*

~99.7% of scores within +/- 3 *SD* of *M*

Areas under the normal curve



Non-normal distributions

Types of non-normal distribution

- Modality
  - Uni-model (one peak)
  - Bi-modal (two peaks)
  - Multi-modal (more than two peaks)
- Skewness
  - Positive (tail to right)
  - Negative (tail to left)
- Kurtosis
  - Platykurtic (Flat)
  - Leptokurtic (Peaked)

Non-normal distributions

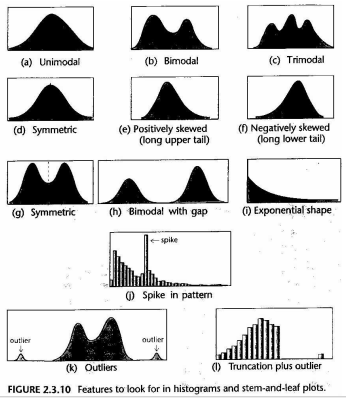
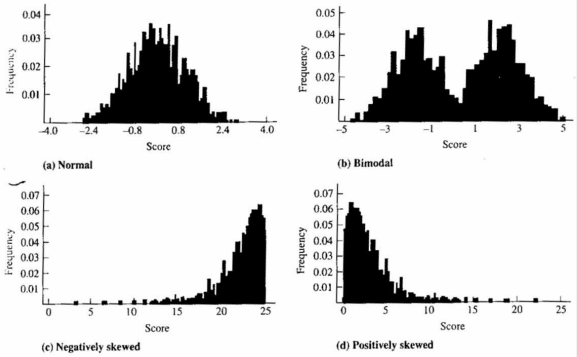
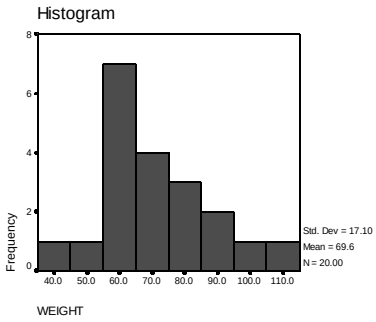


FIGURE 2.3.10 Features to look for in histograms and stem-and-leaf plots.

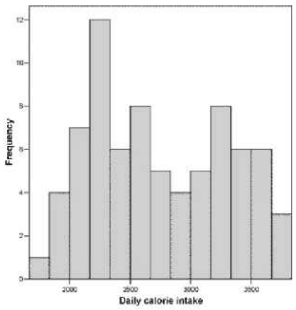
Non-normal distributions



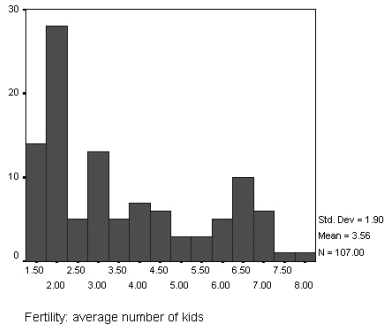
Histogram of weight

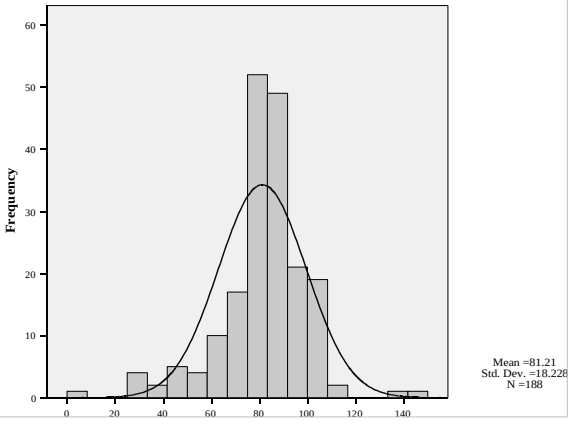


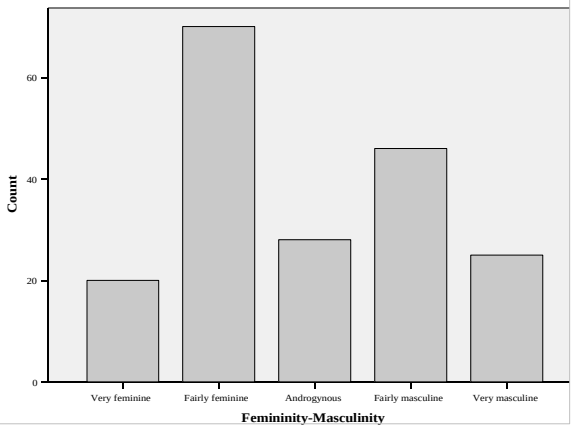
Histogram of daily calorie intake

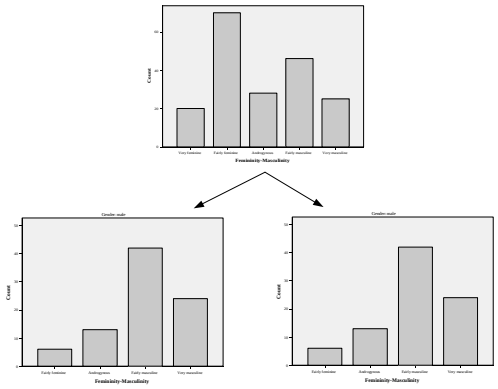


Histogram of fertility









56

Non-normal distribution: Descriptive statistics

- Min & Max
- Range = Max-Min
- Percentiles
- Quartiles
  - –Q1
  - –Mdn = Q2
  - –Q3
  - –IQR = Q3-Q1

57

Effects of skew on measures of central tendency

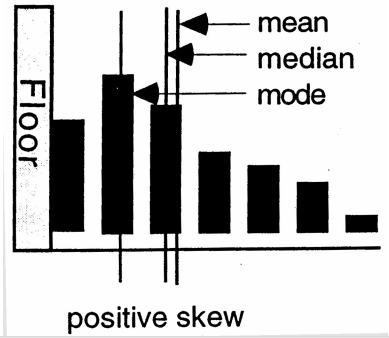
+vely skewed
mode < median < mean

Symmetrical (normal)
mean = median = mode

-vely skewed
mean < median < mode

58

Effects of skew on measures of central tendency



59

Transformations

- Converts data using various formulae
- To achieve normality and allow more powerful tests
- Loses original metric
- Complicates interpretation

60

Graphical techniques

Visualisation

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

Graphs
(Edward Tufte)

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

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 Graphs or via Analyses

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

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

Principles of graphing

Graphical display principles

- Have a clear purpose in mind
- Maximise clarity of information conveyed; minimise clutter
- Find creative, effective ways to show the data
- Substance > fanciness
- Avoid distortions of data
- Clear labelling

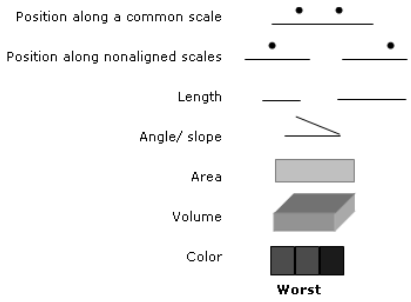
Tufte's graphing guidelines

- 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

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

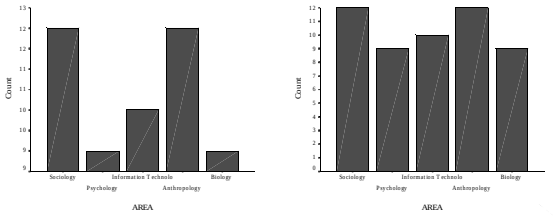
Univariate graphs

Univariate graphs

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

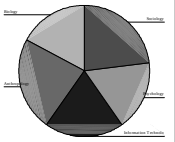
Bar chart (Bar graph)

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



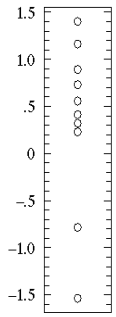
Pie chart

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

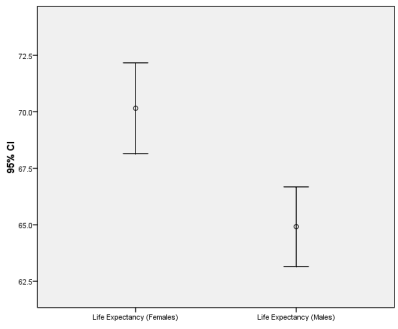


Data plot & error bar

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	01122344455556677778899
	1	01112223333344455555566666666677788899
10 11 11 11 12 12 12 13 13 13 13 14 14 14 15 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
20 20 21 21 22 22 23 23 24 24 24 24 25 25 26 26 27 28 28 29	3	005
30 30 35		

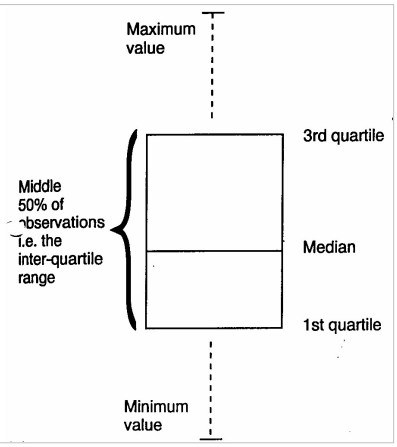
Stem & leaf plot

- Contains actual data
- Collapses tails

```
 Frequency    Stem &  Leaf
     7.00       1 .  &
   192.00       1 .  22223333333
   541.00       1 .  44444444444444455555555555555
   610.00       1 .  66666666666667777777777777777777
   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)
```

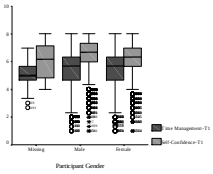
Box plot
(Box & whisker)

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



Box plot

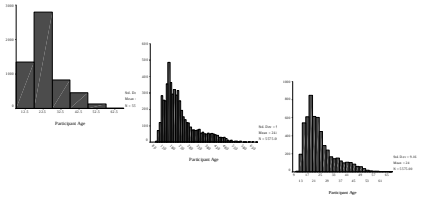
- Alternative to histogram
- Useful for screening
- Useful for comparing variables
- Can get messy - too much info
- Confusing to unfamiliar reader



80

Histogram

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



81

Histogram of male & female heights

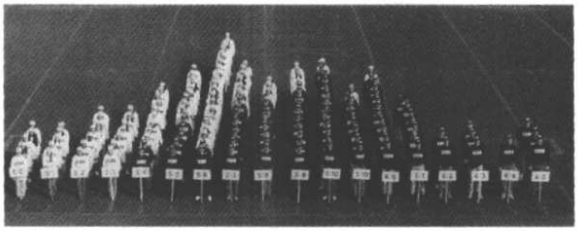
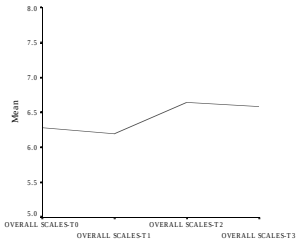


FIGURE 2.3.11 Histogram of heights constructed using the people. Photograph by Peter Morenus in conjunction with Prof. Linda Strausberg, University of Connecticut. Subjects are University of Connecticut genetics students, females in white tops, males in dark tops.

82

Line graph

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

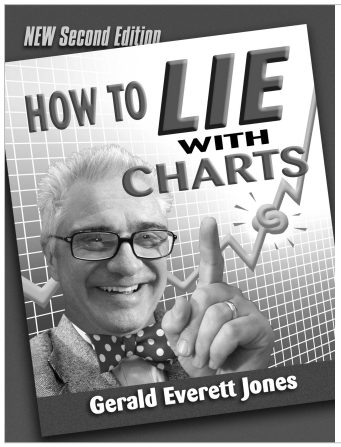


83

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

84



Graphical integrity

(part of academic integrity)

"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

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. Would you expect the mean # of cars owned in Australia to exceed the median?

Review questions

3. Would you expect the mean score on an easy test to exceed the median performance?
4. 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?

Review exercise:
Fill in the cells in 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