Descriptives & Graphing

Image source: http://commons.wikimedia.org/wiki/File:3D_Bar_Graph_Meeting.jpg

Lecture 3

Survey Research & Design in Psychology

James Neill, 2018

Overview: Descriptives & Graphing

1. Getting to know data
2. LOM & types of statistics
3. Descriptive statistics
4. Normal distribution
5. Non-normal distributions
6. Effect of skew on central tendency
7. Principles of graphing
8. Univariate graphical techniques

2

Readings

Howitt & Cramer (2014):

- Chapter 01 - Why statistics?
- Chapter 02 - Some basics: Variability and measurement
- Chapter 03 - Describing variables: Tables and diagrams
- Chapter 04 - Describing variables numerically: Averages, variation and spread
- Chapter 05 - Shapes of distributions of scores
- Chapter 06 - Standard deviation and *z*-scores: The standard unit of measurement in statistics

3

Getting to know data

(how to approach data)

4

Getting to know data

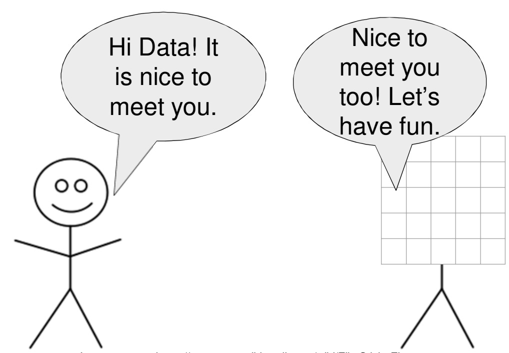


Image source: https://commons.wikimedia.org/wiki/File:Stick_Figure.svg

5

Image source: http://www.flickr.com/photos/analytik/1356366068/


Don't be afraid - you can't break data!
Image source: http://www.flickr.com/photos/rnddave/5094020069

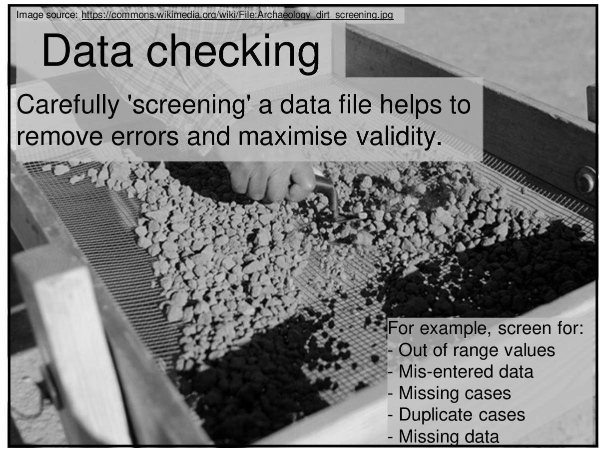
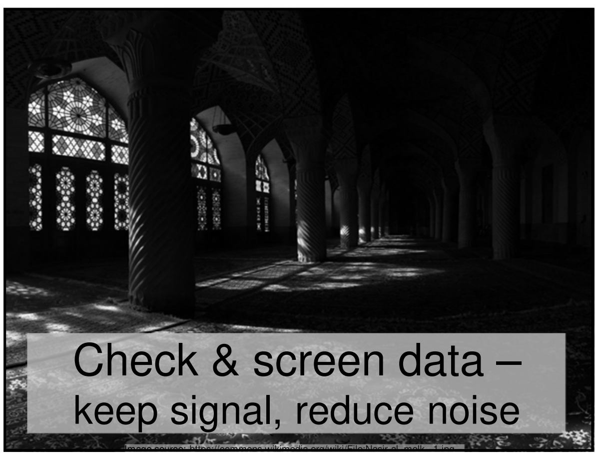


Image source: https://commons.wikimedia.org/wiki/File:Archaeology_dirt_screening.jpg
Data checking
Carefully 'screening' a data file helps to remove errors and maximise validity.
For example, screen for:
- Out of range values
- Mis-entered data
- Missing cases
- Duplicate cases
- Missing data


Check & screen data – keep signal, reduce noise


Explore data

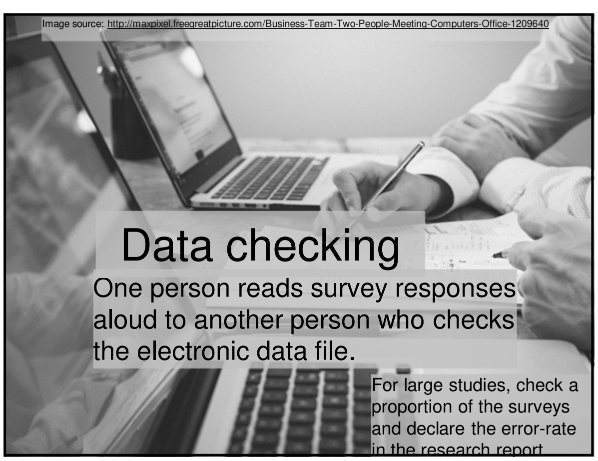


Image source: http://maxpixel.freegreatpicture.com/Business-Team-Two-People-Meeting-Computers-Office-1209640
Data checking
One person reads survey responses aloud to another person who checks the electronic data file.
For large studies, check a proportion of the surveys and declare the error-rate in the research report

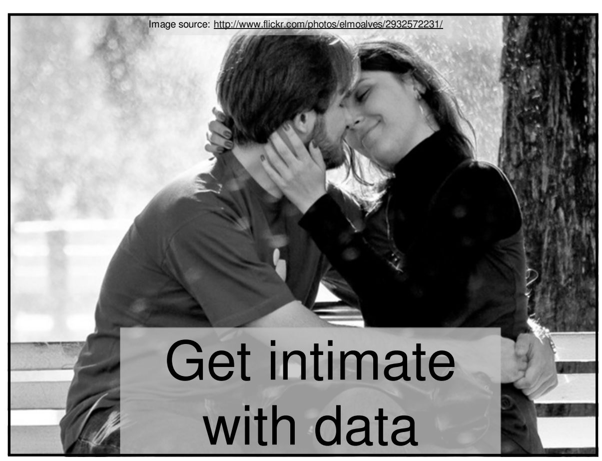
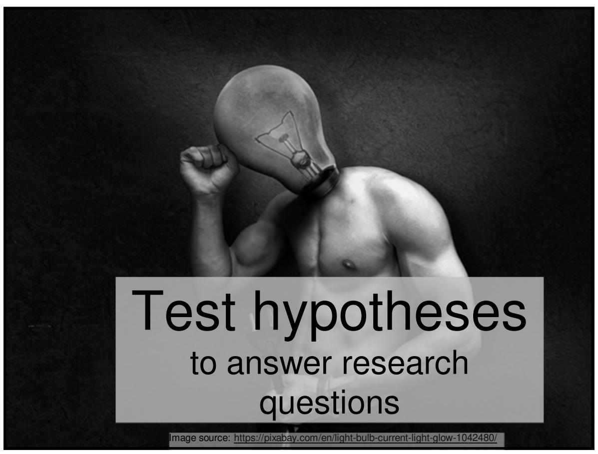


Image source: http://www.flickr.com/photos/elmoalves/2932572231/
Get intimate with data



Level of measurement & types of statistics

Image source: http://www.flickr.com/photos/peanutlen/2228077524/

15

LOM → statistics

Level of measurement determines the type of statistics that can be used, including types of:

- descriptive statistics
- graphs
- inferential statistics

16

LOM - Parametric vs. non-parametric

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

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

DVs = dependent variables

17

Parametric statistics

- Statistics which estimate **parameters** of a population, based on the **normal distribution**
  - **Univariate**: mean, standard deviation, skewness, kurtosis
  - **Bivariate:** correlation, linear regression, *t*-tests
  - **Multivariate**: multiple linear regression, ANOVAs

18

Parametric statistics

- More powerful
  (more sensitive)
- More assumptions
  (population is normally distributed)
- Vulnerable to violations of assumptions
  (less robust)

19

Non-parametric statistics

- Statistics which do not assume sampling from a population which is **normally distributed**
  - There are non-parametric alternatives for many parametric statistics
  - e.g., sign test, chi-squared, Mann-Whitney U test, Wilcoxon matched-pairs signed-ranks test.

20

Non-parametric statistics

- Less powerful
  (less sensitive)
- Fewer assumptions
  (do not assume a normal distribution)
- Less vulnerable to assumption violation
  (more robust)

21

Summary: LOM & statistics

- If a normal distribution can be assumed, use parametric statistics (more powerful)
- If not, use non-parametric statistics (less power, but less sensitive to violations of assumptions)

22

Univariate descriptive statistics

23

Number of variables

Univariate
= one variable

mean, median, mode, histogram, bar chart

Bivariate
= two variables

correlation, *t*-test, scatterplot, clustered bar chart

Multivariate
= more than two variables

reliability analysis, factor analysis, multiple linear regression

24

What to describe?

- **Central tendency**(ies): e.g., frequencies, mode, median, mean
- **Distribution**:
  - **Spread** (dispersion): min., max., range, IQR, percentiles, variance, standard deviation
  - **Shape**: e.g., skewness, kurtosi

25

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.

26

Central tendency

	Mode / Freq. /%s	Median	Mean
Nominal	√	x	x
Ordinal	√	If meaningful	x
Interval	√	√	√
Ratio	If meaningful	√	√

27

Distribution

- Measures of shape, spread, dispersion, and deviation from the central tendency

Non-parametric:
- Min. and max.
- Range
- Percentiles

Parametric:
- *SD*
- Skewness
- Kurtosis

28

Distribution

	Min / Max, Range	Percentile	*Var / SD*
Nominal	x	x	x
Ordinal	√	If meaningful	x
Interval	√	√	√
Ratio	√	√	√

29

Descripives for nominal data

- **Nominal LOM** = 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)

30

Descripives for ordinal data

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

31

Descripives for interval data

- **Interval LOM** = 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)

32

Descriptives for ratio data

- **Ratio** = Numbers convey order and distance, meaningful 0 point
- As for interval, use median, mean, *SD*, skewness etc.
- Can also use ratios (e.g., Group A is twice as "large" as Group B)

33

Mode (*Mo*)

- Most common score - highest point in a frequency distribution – a real score – the most common response
- Suitable for all levels of data, but may not be appropriate for ratio (continuous)
- Not affected by outliers
- Check frequencies and bar graph to see whether it is an accurate and useful statistic

34

Frequencies (*f*) and percentages (%)

- # of responses in each category
- % of responses in each category
- Frequency table
- Visualise using a bar or pie chart

35

Median (*Mdn*)

- Mid-point of distribution (Quartile 2, 50th percentile)
- Not badly affected by outliers
- May not represent the central tendency in skewed data
- If Median is useful, other percentiles may also be worth reporting

36

Summary: Descriptive statistics

- **Level of measurement** and **normality** determines whether data can be treated as **parametric**
- Describe the **central tendency**
  - –Frequencies, Percentages
  - –Mode, Median, Mean
- Describe the **distribution**:
  - –Min., Max., Range, Quartiles
  - –Standard Deviation, Variance

37

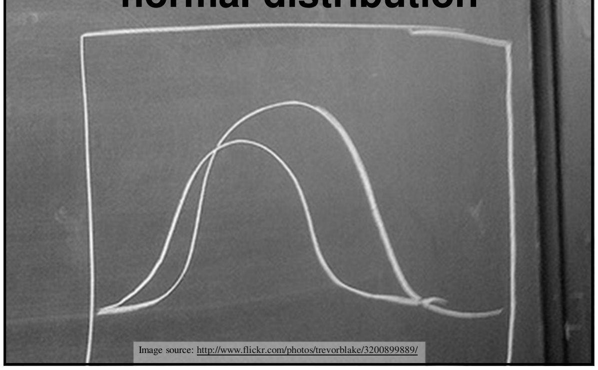
Four moments of a normal distribution

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

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

40

Properties of the normal distribution



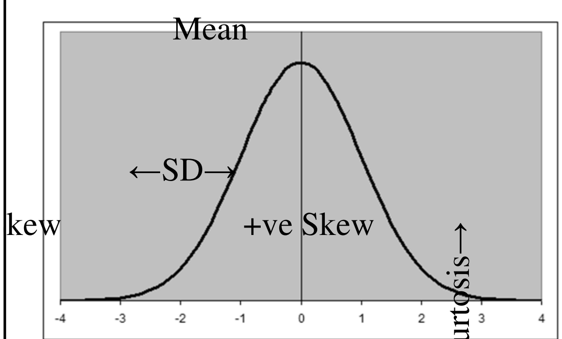
Mean (1st moment)

- Average score

  Mean = Σ X / N
- Use for normally distributed ratio or interval (if treating as continuous) data.
- Influenced by extreme scores (outliers)

41

Four moments of a normal distribution



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* = 1.999).

42

Standard deviation (2nd moment)

- *SD* = square root of the variance $= \frac{\Sigma (X - \bar{X})^2}{N - 1}$
- Use for normally distributed interval or ratio data
- Affected by outliers
- Can also derive Standard Error (SE) = *SD* / square root of *N*

43

Skewness (3rd moment)

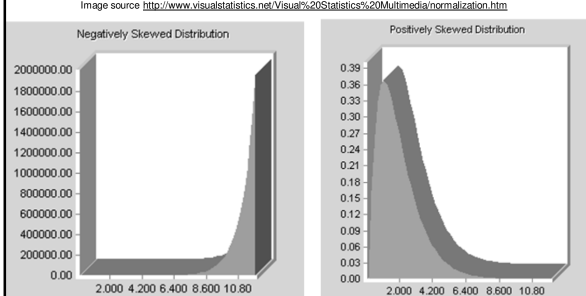
- Lean of distribution
  - +ve = tail to right
  - -ve = tail to left
- Skew be caused by an outlier, or ceiling or floor effects
- Skew be accurate
  (e.g., cars owned per person would have a skewed distribution)

44

Skewness (3rd moment)

(with ceiling and floor effects)

Image source http://www.visualstatistics.net/Visual%20Statistics%20Multimedia/normalization.htm



- Negative skew
- Ceiling effect

- Positive skew
- Floor effect

Kurtosis (4th moment)

- Flatness vs. peakedness of distribution:
  - +ve = peaked
  - -ve = flattened
- Altering the X &/or Y axis can artificially make a distribution look more peaked or flat – add a normal curve to help judge kurtosis visually.

46

Kurtosis (4th moment)

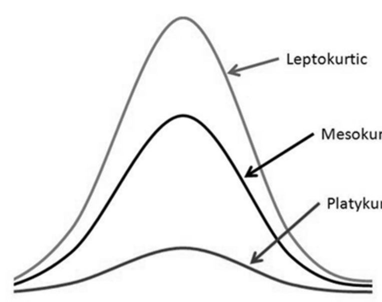


Image source: https://classconnection.s3.amazonaws.com/65/flashcards/2185065/jpg/kurtosis-142C1127AF2178FB244.jpg

Severity of skewness and kurtosis

- View histogram with normal curve
- Deal with outliers
- Rule of thumb:
  Skewness and kurtosis > -1 or < 1 is generally considered to sufficiently normal for meeting the assumptions of parametric inferential statistics
- Significance tests of skewness:
  Tend to be overly sensitive
  (therefore avoid using)

48

Areas under the normal curve

If distribution is normal (bell-shaped):

~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*

49

Non-normal distributions

- Modality
  - Uni-modal (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)

52

Areas under the normal curve

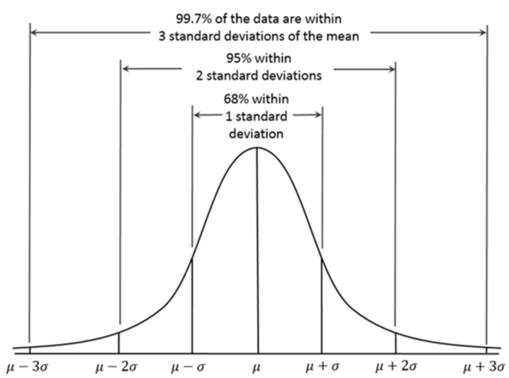


Image source: https://commons.wikimedia.org/wiki/File:Empirical_Rule.PNG

Non-normal distributions

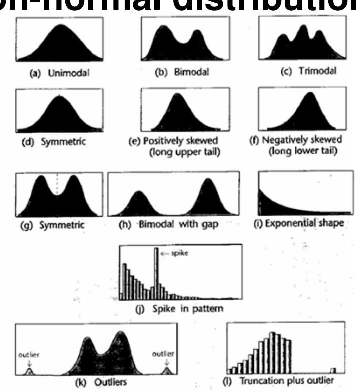
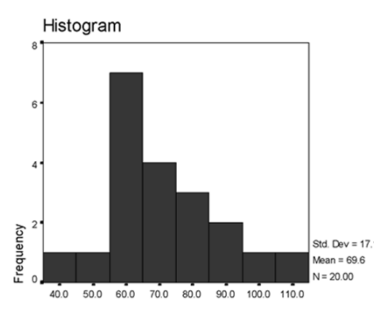


FIGURE 2.3.10 Features to look for in histograms and stem-and-leaf plots.

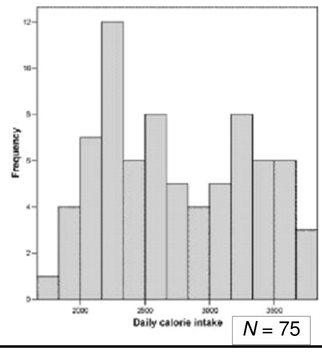
Non-normal distributions

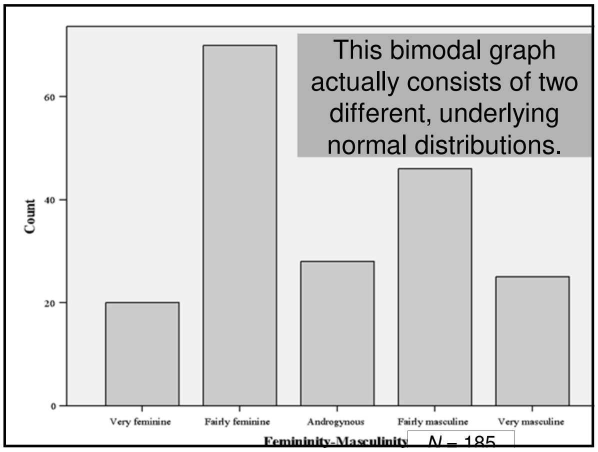
51

Histogram of people's weight

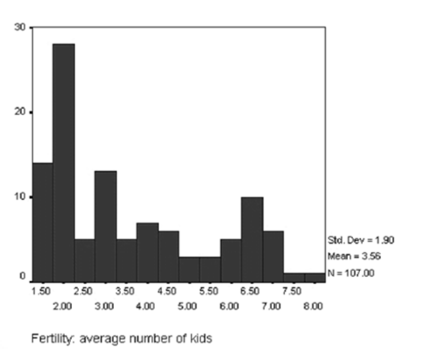


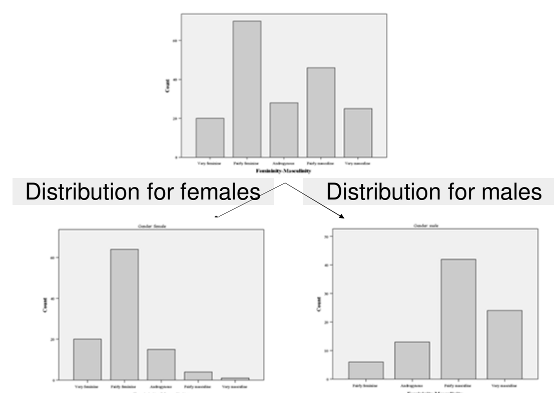
Histogram of daily calorie intake



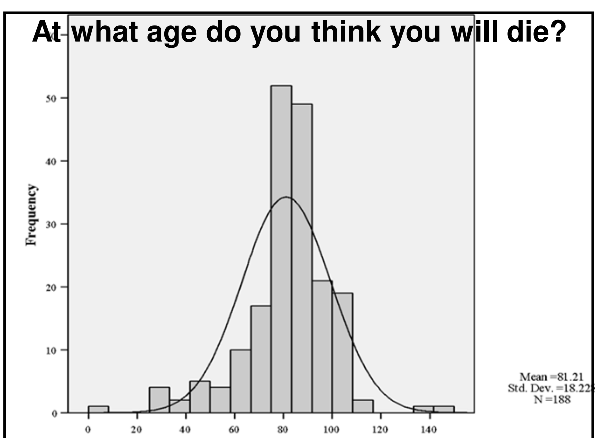


Histogram of fertility





At what age do you think you will die?



Non-normal distribution:

Use non-parametric descriptive statistics

- Min. & Max.
- Range = Max. - Min.
- Percentiles
- Quartiles
  - –Q1
  - –Median (Q2)
  - –Q3
  - –IQR (Q3-Q1)

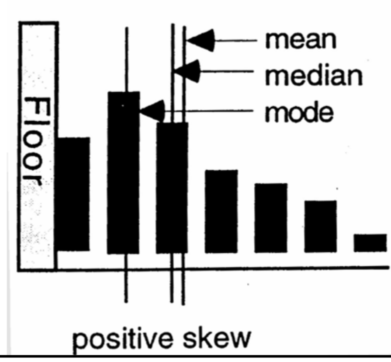
60

Effects of skew on measures of central tendency

+vely skewed distributions
mode < median < mean
symmetrical (normal) distributions
mean = median = mode
-vely skewed distributions
mean < median < mode

61

Effects of skew on measures of central tendency



Transformations

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

63

Review questions

1. If a survey question produces a "floor effect", where will the mean, median and mode lie in relation to one another?

64

Review questions

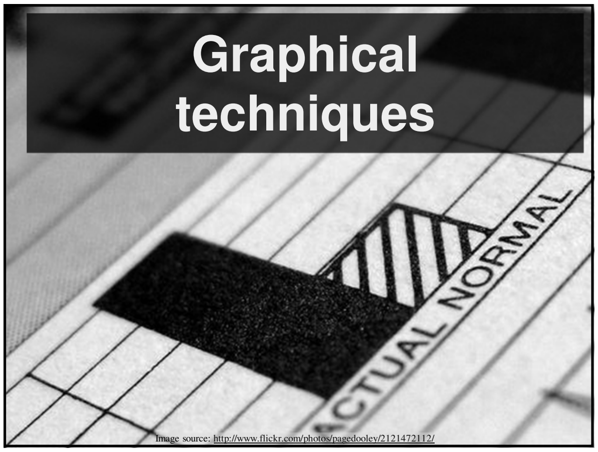
2. Would the mean # of cars owned in Australia exceed the median?

65

Review questions

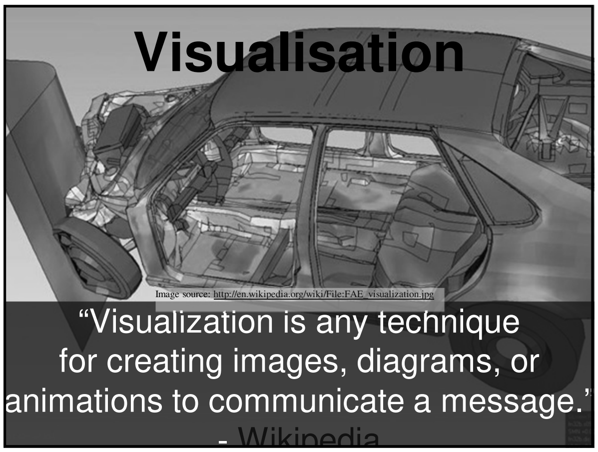
3. Would the mean score on an easy test exceed the median performance?

66



Graphical techniques

Image source: http://www.flickr.com/photos/pagedooley/2121472112/



Visualisation

Image source: http://en.wikipedia.org/wiki/File:FAE_visualization.jpg

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

Science is beautiful

(Nature Video)

Image source::http://commons.wikimedia.org/wiki/File:Parodyfilm.png

(Youtube – 5:30 mins)

69

Is Pivot a turning point for web exploration?

(Gary Flake)

Image source:http://commons.wikimedia.org/wiki/File:Parodyfilm.png

(TED talk - 6 min.)

70

Principles of graphing

- Clear purpose
- Maximise clarity
- Minimise clutter
- Allow visual comparison

71

Graphs (Tufte)

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

72

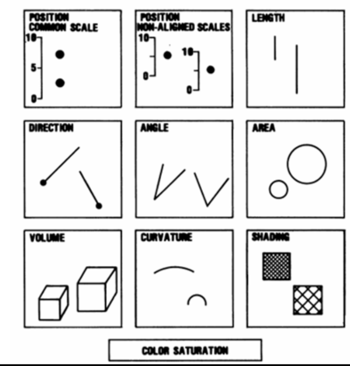
Graphing steps

1 Identify purpose of the graph (make large amounts of data coherent; present many #s in small space; encourage the eye to make comparisons)
2 Select type of graph to use
3 Draw and modify graph to be clear, non-distorting, and well-labelled (maximise clarity, minimise clarity; show the data; avoid distortion; reveal data at several levels/layers)

73

Cleveland's hierarchy

Image source:https://priceonomics.com/how-william-cleveland-turned-data-visualization/



Graphing software

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

74

Univariate graphs

- Bar graph
- Pie chart
- Histogram
- Stem & leaf plot
- Data plot / Error bar
- Box plot

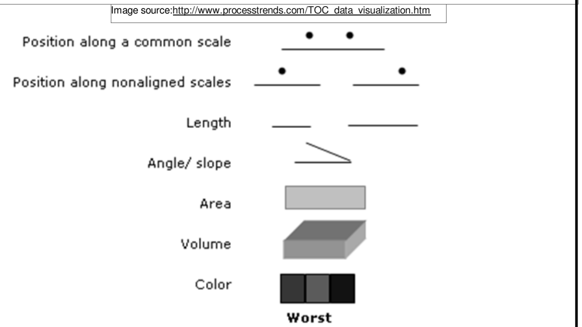
Non-parametric
i.e., nominal or ordinal

Parametric
i.e., normally distributed interval or ratio

77

Cleveland's hierarchy

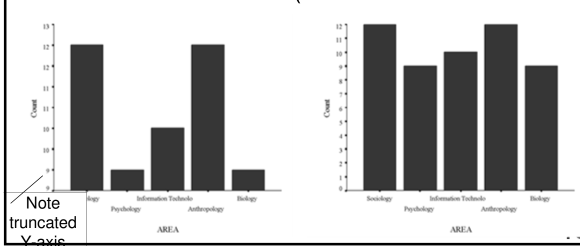
Image source:http://www.processtrends.com/TOC_data_visualization.htm



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

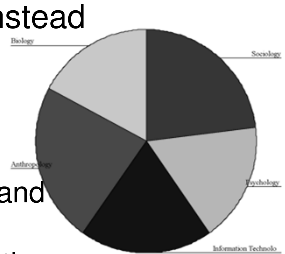
Bar chart

- Allows comparison of heights of bars
- X-axis: Collapse if too many categories
- Y-axis: Count/Frequency or % - truncation exaggerates differences
- Can add data labels (data values for each bar)



Pie chart

- Use a bar chart instead
- Hard to read
  - Difficult to show
    - Small values
    - Small differences
  - Rotation of chart and position of slices influences perception



79

Histogram of male and female heights

Image source: Wild, C. J., & Seber, G. A. F. (2000). *Chance encounters: A first course in data analysis and inference.* New York: Wiley

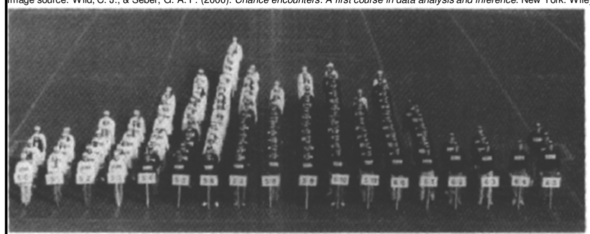


FIGURE 2.3.11 Histogram of heights constructed using the people. Photograph by Peter Morenus in conjunction with Prof. Linda Strausberg, University of Connecticut. Subjects are University of Connecticut genetics students, females in white tops, males in dark tops.

Wild & Seber (2000)

Pie chart → Use bar chart instead

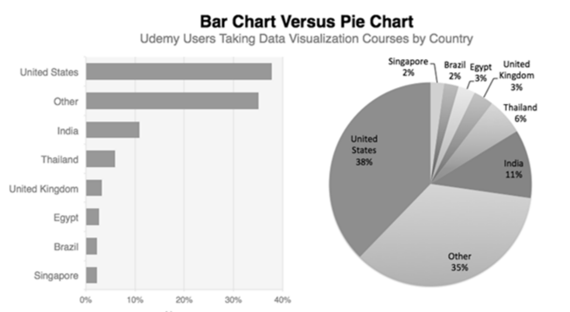


Image source: https://priceonomics.com/how-william-cleveland-turned-data-visualization/

80

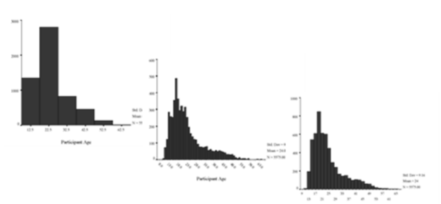
Stem and leaf plots

- Use for ordinal, interval and ratio data (if rounded)
- 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	0112234445556677778899
10 11 11 11 12 12 12 13 13 13 13 13 14 14 14 15 15 15 15 15 15 16 16 16 16 16 16 16 16 16 16 17 17 17 18 18 18 18 19 19	1	0111222333334445555556666666666777888899
20 20 21 21 22 22 23 23 24 24 24 24 25 25 26 26 27 28 28 29	2	00112233444455667889
30 30 35	3	005

Histogram

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



81

Stem and leaf plot

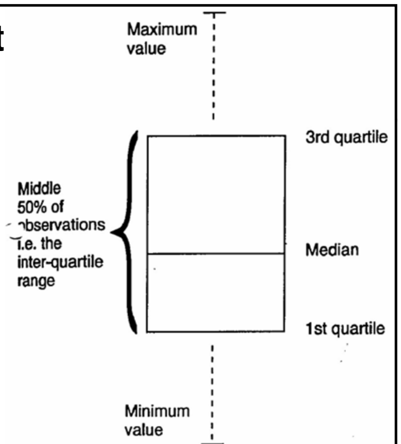
- Contains actual data
- Collapses tails
- Underused alternative to histogram

```
Frequency    Stem &  Leaf
     7.00       1 .  &
   192.00       1 .  22223333333
   541.00       1 .  4444444444444445555555555555555
   610.00       1 .  66666666666666777777777777777777777
   849.00       1 .  888888888888888888888889999999999999999999999999
   614.00       2 .  00000000000000000111111111111111111
   602.00       2 .  2222222222222222333333333333333333
   447.00       2 .  4444444444444455555555555
   291.00       2 .  66666666667777777
   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)
```

84

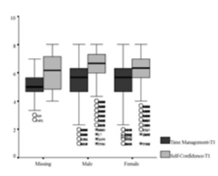
Box plot
(Box & whisker)

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



Box plot (Box & whisker)

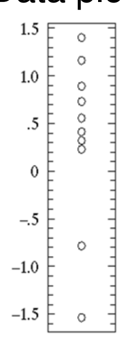
- Alternative to histogram
- Useful for screening
- Useful for comparing variables
- Can get messy - too much info
- Confusing to unfamiliar reader



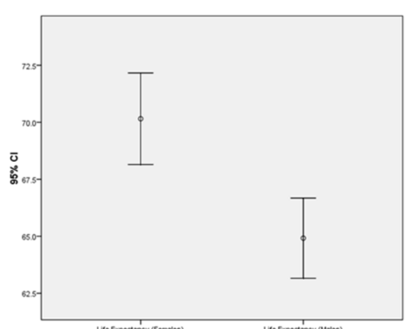
86

Data plot & Error bar

Data plot

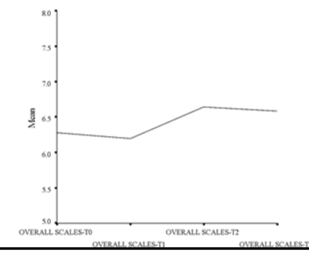


Error bar

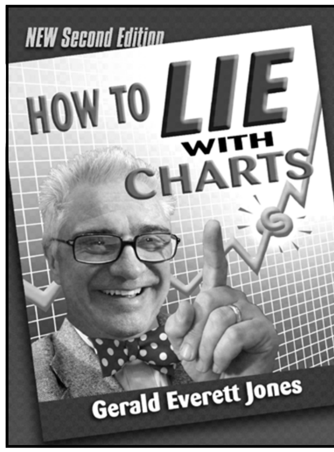


Line graph

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



88



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

90

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

91

Review exercise:
Fill in the cells in this table

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

Answers: http://goo.gl/Ln9e1

92

References

1 Chambers, J., Cleveland, B., Kleiner, B., & Tukey, P. (1983). *Graphical methods for data analysis*. Boston, MA: Duxbury Press.
2 Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.
3 Jones, G. E. (2006). *How to lie with charts*. Santa Monica, CA: LaPuerta.
4 Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
5 Tufte. E. R. (2001). *Visualizing quantitative data*. Cheshire, CT: Graphics Press.
6 Tukey J. (1977). *Exploratory data analysis*. Addison-Wesley.
7 Wild, C. J., & Seber, G. A. F. (2000). *Chance encounters: A first course in data analysis and inference*. New York: Wiley.

93

Next lecture

Correlation
- Covariation
- Purpose of correlation
- Linear correlation
- Types of correlation
- Interpreting correlation
- Assumptions / limitations

94