Descriptives & Graphing

Lecture 3

Survey Research & Design in Psychology

James Neill, 2016

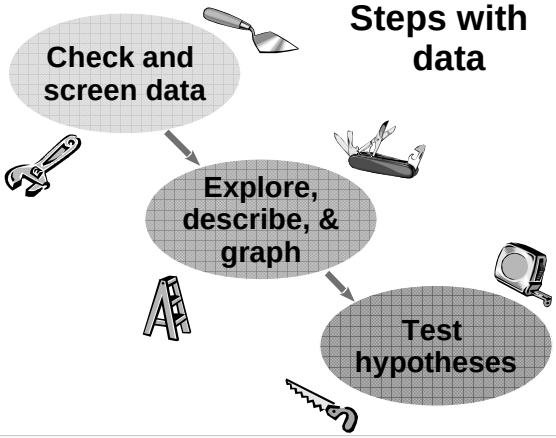
Overview: Descriptives & Graphing

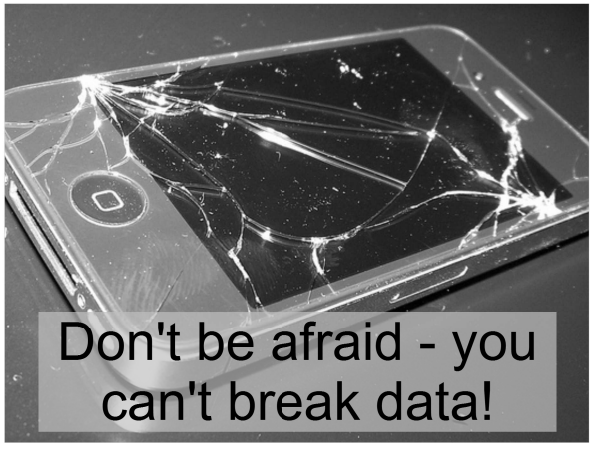
1. Steps with data
2. Level of measurement & types of statistics
3. Descriptive statistics
4. Normal distribution
5. Non-normal distributions
6. Effect of skew on central tendency
7. Principles of graphing
8. Univariate graphical techniques

Steps with data

(how to approach data)

Steps with data





Data checking

- Have one one person read the 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.

Data screening

- Carefully 'screening' a data file helps to minimise errors and maximise validity.
- For example, screen for:
  - Out of range values (min. and max.)
  - Mis-entered data
  - Missing cases
  - Duplicate cases
  - Missing data

Level of measurement & types of statistics

Golden rule of data analysis

Level of measurement determines type of descriptive statistics and graphs

Level of measurement 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**
- Key parametric statistics:
  - Univariate: *M*, *SD*, skewness, kurtosis → *t*-tests, ANOVAs
  - Bi/multivariate: *r* → linear regression, multiple linear regression

Parametric statistics

- More powerful (more sensitive)
- More assumptions (normal distribution)
- More vulnerable to violations of assumptions

Non-parametric statistics
(Distribution-free tests)

- Fewer assumptions (do not assume a normal distribution)
- Common non-parametric statistics:
  - Frequency → sign test, chi-squared
  - Rank order → Mann-Whitney U test, Wilcoxon matched-pairs signed-ranks test

Univariate descriptive statistics

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

What do we want to *describe*?

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

Measures of distribution

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

Non-parametric:

- Min and max
- Range
- Percentiles

Parametric:

- *SD*
- Skewness
- Kurtosis

Measures of spread / dispersion / deviation

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

Descriptives 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)

Descriptives 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

Descriptives 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)

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., Category A is twice as large as Category B)

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

Frequencies (*f*) and percentages (%)

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

Median (*Mdn*)

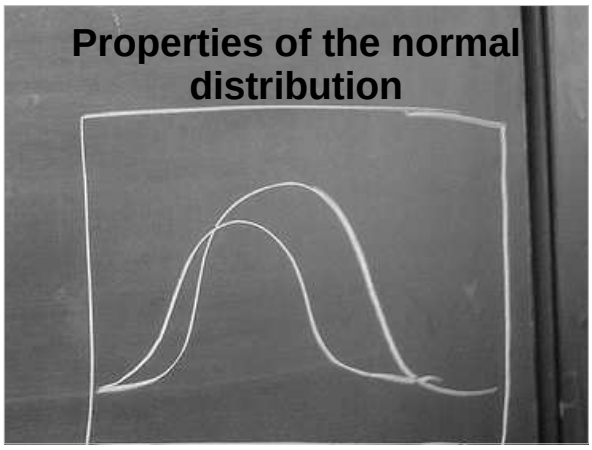
- Mid-point of distribution (Quartile 2, 50th percentile)
- Not badly affected by outliers
- May not represent the central tendency in skewed data
- If the Median is useful, then consider what other percentiles may also be worth reporting

Summary: Descriptive statistics

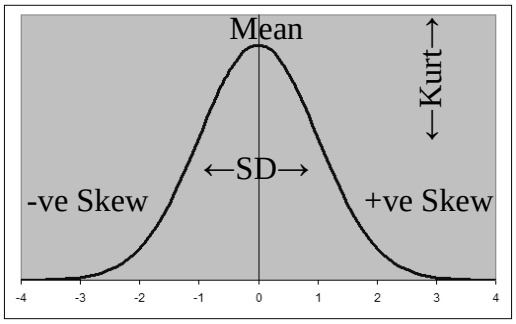
- **Level of measurement** and **normality** determines whether data can be treated as **parametric**
- Describe the **central tendency**
  - Frequencies, Percentages
  - Mode, Median, Mean
- Describe the **variability**:
  - Min, Max, Range, Quartiles
  - Standard Deviation, Variance

31

Properties of the normal distribution



Four moments of a normal distribution



33

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)

34

Mean (1st moment)

- Average score

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

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

35

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).

36

Standard deviation (2nd moment)

- *SD* = square root of the variance

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

- For normally distributed interval or ratio data
- Affected by outliers
- Can also derive the Standard Error (SE) = *SD* / square root of *N*

37

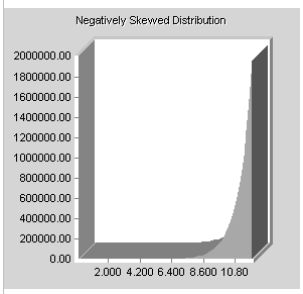
Skewness (3rd moment)

- 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., cars owned per person would have a skewed distribution)

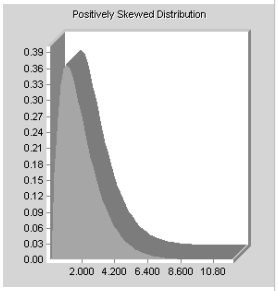
38

Skewness (3rd moment)

(with ceiling and floor effects)



- Negative skew
- Ceiling effect



- Positive skew
- Floor effect

39

Kurtosis (4th moment)

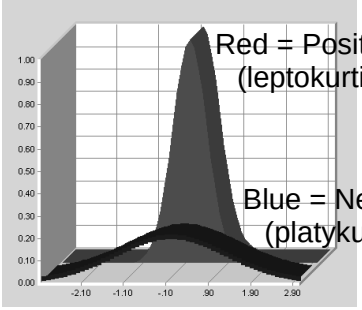
- Flatness or peakedness of distribution

  +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.

40

Kurtosis (4th moment)



41

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 sufficiently normal for meeting the assumptions of parametric inferential statistics
- Significance tests of skewness:
  Tend to be overly sensitive
  (therefore avoid using)

42

Areas under the normal curve

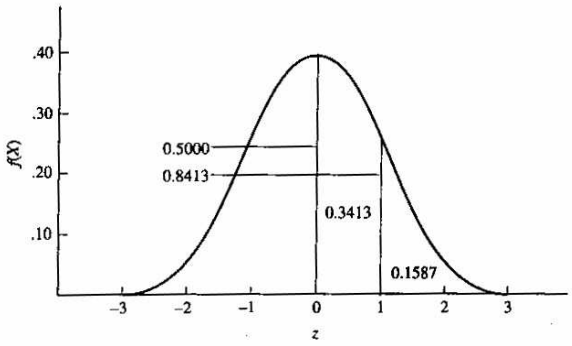
If distribution is normal
(bell-shaped - or close):

~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-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)

Non-normal distributions

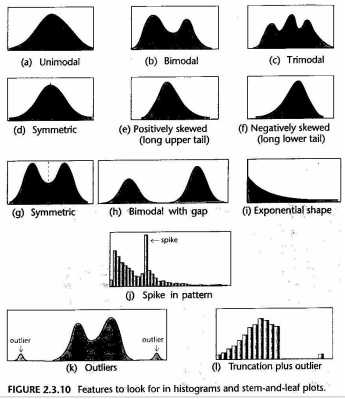
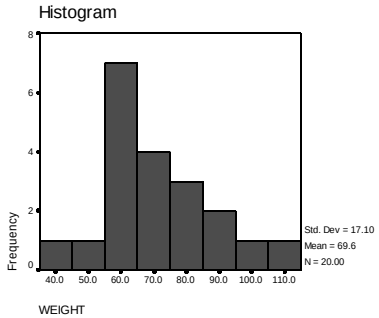
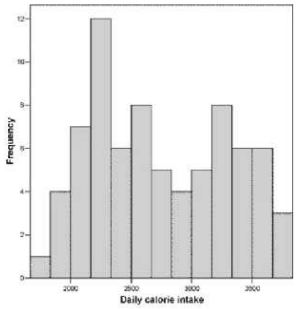


FIGURE 2.3.10 Features to look for in histograms and stem-and-leaf plots.

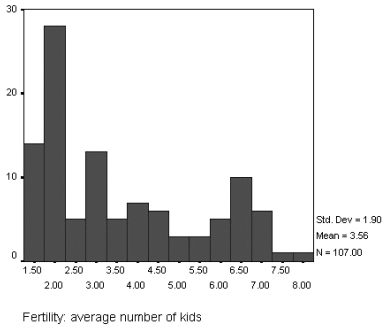
Histogram of weight

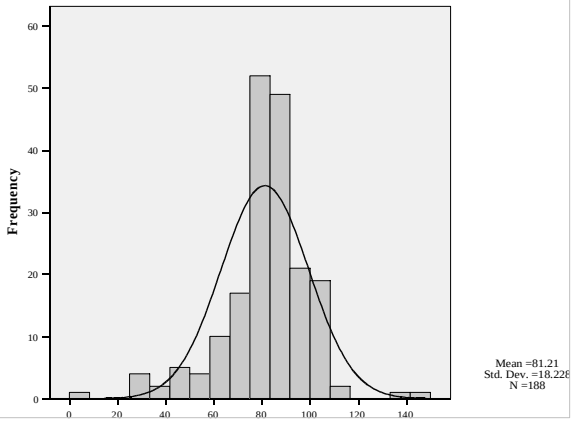


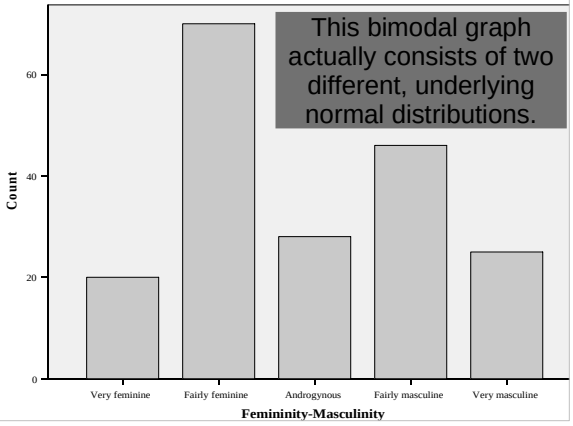
Histogram of daily calorie intake

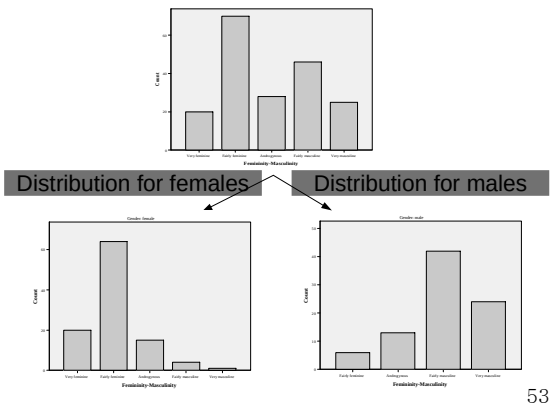


Histogram of fertility









Non-normal distribution:
Use non-parametric descriptive statistics

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

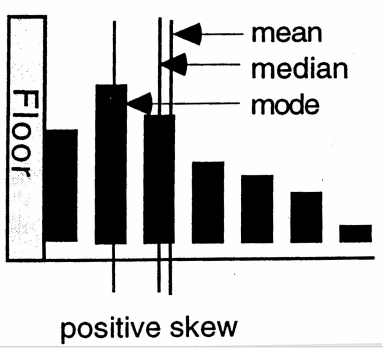
Effects of skew on measures of central tendency

+vely skewed distributions
mode < median < mean

symmetrical (normal) distributions
mean = median = median

-vely skewed distributions
mean < median < mode

Effects of skew on measures of central tendency



positive skew

Transformations

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

Review questions

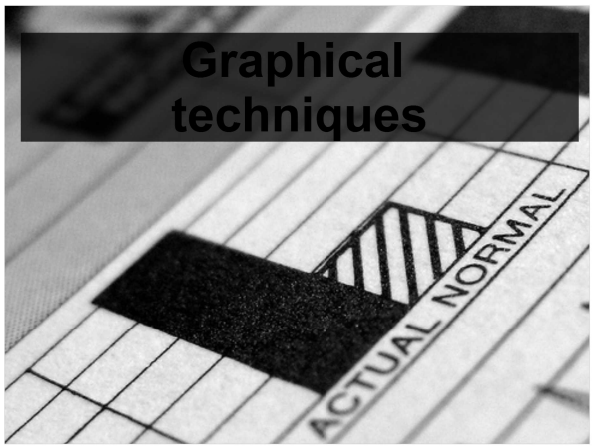
1.If a survey question produces a 'floor effect', where will the mean, median and mode lie in relation to one another?

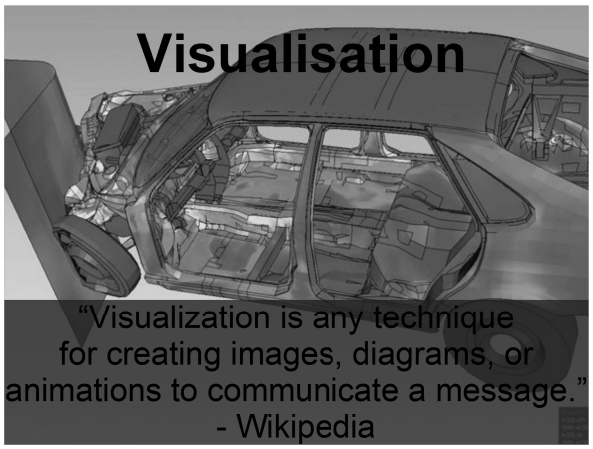
Review questions

2.Would 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?





Science is beautiful
(Nature Video)

(Youtube – 5:30 mins)

63

Is Pivot a turning point for web exploration?
(Gary Flake)

(TED talk - 6 min.)

64

Principles of graphing

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

65

Graphs
(Edward Tufte)

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

66

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)

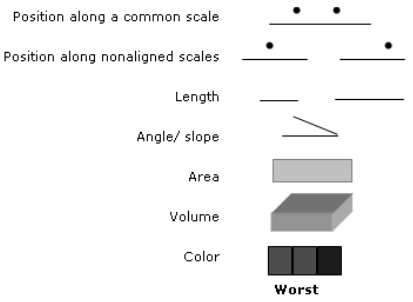
Software for data visualisation (graphing)

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

Cleveland's hierarchy



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

Univariate graphs

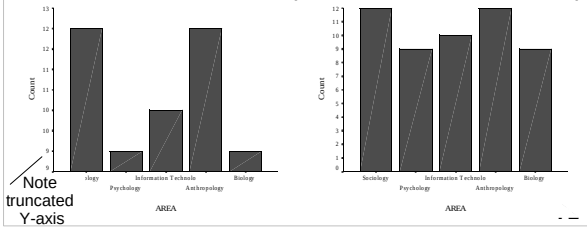
- Bar graph
- Pie chart
- Histogram
- Stem & leaf plot
- Data plot / Error bar
- Box plot

Non-parametric
i.e., nominal, ordinal, or non-normal interval or ratio

Parametric
i.e., normally distributed interval or ratio

Bar chart (Bar graph)

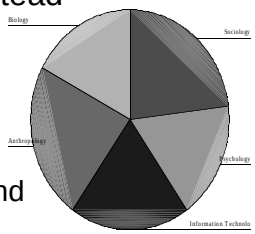
- Allows comparison of heights of bars
- X-axis: Collapse if too many categories
- Y-axis: Count/Frequency or % - truncation exaggerates differences
- Can add data labels (data values for each bar)



Note truncated Y-axis

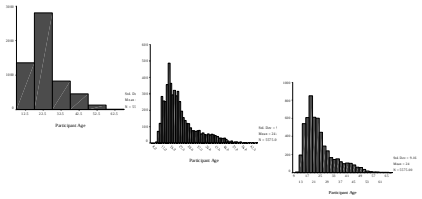
Pie chart

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



Histogram

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



Histogram of male & female heights

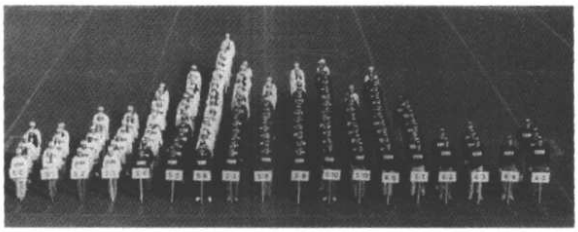


FIGURE 2.3.11 Histogram of heights constructed using the people. Photograph by Peter Morenus in conjunction with Prof. Linda Strausberg, University of Connecticut. Subjects are University of Connecticut genetics students, females in white tops, males in dark tops.

Wild & Seber (2000)

Stem & leaf plot

- 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	01122344455566777788899
	1	011122233333444555556666666667778888899
10 11 11 11 12 12 12 13 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
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		

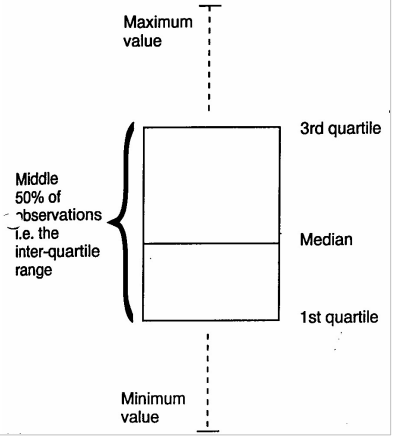
Stem & leaf plot

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

```
Frequency    Stem &  Leaf
    7.00        1 .  &
  192.00        1 .  22223333333
  541.00        1 .  44444444444444455555555555555
  610.00        1 .  66666666666666777777777777777777
  849.00        1 .  88888888888888888888888888999999999999999999
  614.00        2 .  00000000000000001111111111111111
  602.00        2 .  22222222222222233333333333333
  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)
```

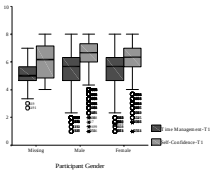

Box plot (Box & whisker)

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



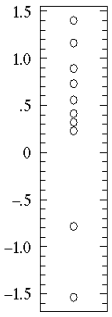
Box plot (Box & whisker)

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

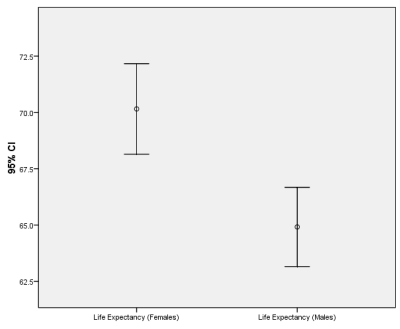


Data plot & error bar

Data plot

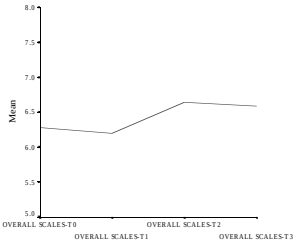


Error bar

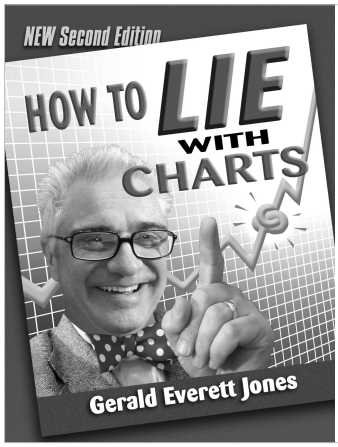


Line graph

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



80



Graphical integrity

(part of academic integrity)

81

"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

82

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

83

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

84

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.

Open Office Impress

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