Contents

Contributors:
- Daniel Fitz (Sanchez)

0.1 Assessment

- Two Projects
  - Visualization
  - Graphics
- Each has 3 components
  - Proposal (pass/fail)
  - Presentation (inadequate/poor/good/excellent)
  - Report (1-7)

For the visualization project, show that you can analysis, understand, and/or communicate or teach about data

- Multiple independent variables
- Multiple dependent variables
- Complex behavior over space
- Complex behavior over time

Chapter 1

Lecture Notes

1.1 Data Visualization

The use of images to provide insight into phenomena. Should reveal data:

- show the data, honestly
- thought-provoking (not distracting)
- efficient (many data in little space)
- encourage comparison
- expose comparison
- serve a purpose
- link closely to descriptive statistics/text

1.1.1 Visualisation Procedure

Iteractive process:

- Locate/acquire data
- Parse data
- Filter data
- Clean/analyse/derive
- Map to geometry
- Render
- Interact

1.1.2 Data acquisition

Access considerations:

- Need a reliable (credible) source (e.g. govt/university)
- Need the right to use the data
- Acknowledge source
- May need to register/pay
- May have to apply in writing
- Download directly/automatically?
- Dataset[s] may be huge/dynamic
- Can their server cope?
- Be a good internet citizen (... or get blocked)

1.2 Univariate data

Univariate data: multiple measurements for one thing

Bivariate data: multiple measurements of two things, *temperature and windspeed at a station*

Multivariate data: multiple measurements of 3 or more things

1.2.1 Descriptive Statistics

Measures of variation

Ranges: max-min, inter-quartile, boxplots

Standard Deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

Variance: s^2

Skewness: asymmetry $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}$ also (mean - mode)/s

Kurtosis: flatness (platykurtic) or sharpness (leptokurtic)

Types of errors in data:

- human and machine
- recording errors
- transcription/storage errors
- precision and rounding errors
- unit errors
- false presences/absences
- ... and so on

Two kinds of errors affecting all of our data:

Random error: This affects the **precision** of the data

Systematic error: This affects the **accuracy** of the data

1.3 Bivariate data

- **Paired** measurements of two quantitative variables/obervations
- could be just two variables, interested in their **relationship**
- or could be a response (y) to some factor (x)
- can still use univariate methods (quartiles, mean-differences, etc)