Exploring data Rote analysis vs. snooping



Spurious correlations

There's a whole website about this What can you do? The best you can

- Identify scientific questions
- Distinguish between exploratory and confirmatory analysis

- Pre-register studies when possible
- Keep an exploration and analysis journal
- Explore predictors and responses separately at first

1 Individual variables

- Look at location and shape
- Maybe with different sets of grouping variables
- Contrasts
 - Parametric vs. non-parametric
 - Exploratory vs. diagnostic
 - Data vs. inference

Means and standard errors



Means and standard deviations



Means and standard deviations



Bike example



Standard errors



Data shape



Data shape



Data shape and weight



Log scales

- In general:
 - $-\,$ If your logged data span <3 decades, use human-readable numbers (e.g., 10-5000 kilotons per hectare)
 - If not, just embrace "logs" (log10 particles per ul is from 3-8)
 - * But remember these are not physical values
- I love natural logs, but not as axis values
 - Except to represent proportional difference!

2 Bivariate data

Banking

- Banking is a real thing
 - Even though many examples are bogus
- Since the point is make patterns visually clear, trial-and-error is usually as good as algorithm
 - But it is worth considering

Sunspots





Scatter plots

- Depending on how many data points you have, scatter plots may indicate relationships clearly
- They can often be improved with trend interpolations
 - Interpolations may be particularly good for discrete responses (count or true-false)

Scatter plot



Seeing the density better



Maybe fixed



Two loess trend lines



Many loess trend lines



Theory of loess

- Local smoother (locally flat, linear or **quadratic**)
- Neighborhood size given by alpha
 - Points in neighborhood are weighted by distance
- Check help function for loess

Robust methods

- Loess is local, but not robust
 - Uses least squares, can respond strongly to outliers
- R is has a very flexible function called rlm to do robust fitting
 - Not local
 - But can be combined with splines



Fitting comparison

- - Contours
 - use _density_2d() to fit a two-dimensional kernel to the density
 - hexes
 - use geom_hex to plot densities using hexes
 - this can also be done using rectangles for data with more discrete values

Contours



Contours



Hexes

Hexes



Color principles

- Use clear gradients
- If zero has a physical meaning (like density), go in just one direction
 - e.g., white to blue, white to red
 - If the map contrasts with a background, zero should match the background
- If there's a natural *middle*, you can use blue to white to red, or something similar

3 Multiple dimensions

- Three dimensional data is a lot like two-d with densities: contour plots are good
- Pairs plots: pairs, ggpairs

Pairs example



4 Multiple factors

- Use boxplots and violin plots
- Make use of facet_wrap and facetgrid
- Use different combinations (e.g., try plots with the same info, but different factors on the axes vs. in the colors or the facets)

©2018, Jonathan Dushoff and Ben Bolker. May be reproduced and distributed, with this notice, for non-commercial purposes only.