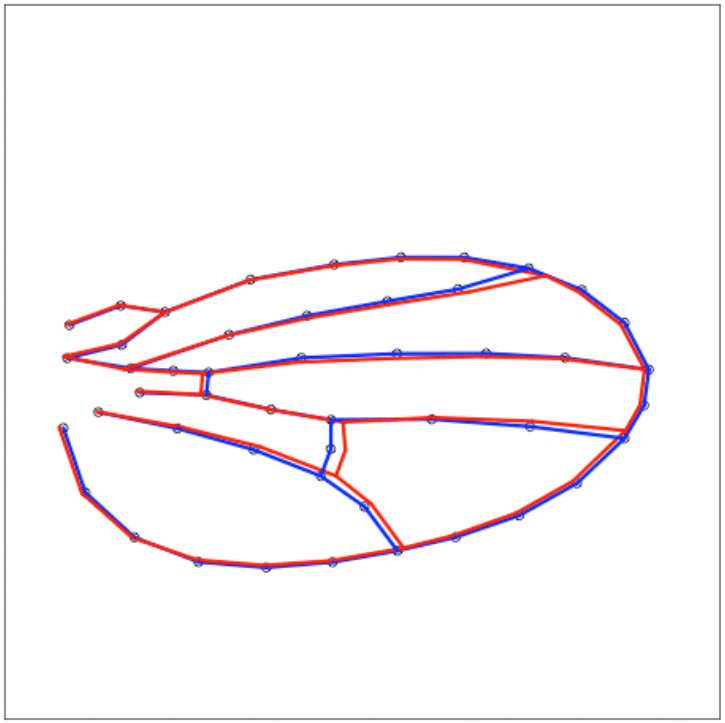
Exploratory Data Analysis

Numerical quantities focus on expected values, graphical summaries on unexpected values.

John Tukey



SMEE ZOL851 Sept 9th 2014

Goals for today

- Think about how we can graphically examine our data.
- Review the basics of exploratory data analysis.
- Go through some R code demonstrating how to get simple plots.

Readings for todays material (see syllabus)

- EMD: Chapter 2 in Bolker.
- Chapter 1 of Dalgaard (R Specific advice for basic plotting).
- Chapter 5 in the “R Book” would be very useful (link is available on ANGEL, in the folder for this lecture).
- Also see Zuur et al paper on ANGEL
- Online lecture on EDA (link on ANGEL)

Ignore!!!
Readings for Sept 13th -18th 2014

- Readings from Crawley (R Book) may be helpful (again for review)
- Crawley Chapter 9, 10 (pages 388-417).
- Crawley pages 449-486, 489-509 (useful for everyone)
- There are also a few manuscripts that may be useful

(Packard & Boardman; McArdle et al.; Garcioa-Berthou; Jackson).

What you need to read and review will depend a great deal on your background.

The Big picture

- Before we can proceed with statistical analyses, we need to examine our data carefully for two major reasons.
- Detect possible errors (measurement, transcription), outliers (real measures, but very extreme).
- While we should have a good idea of the model form before we examine the data, it can help to address specific issues of model specification.

Exploratory Data Analysis

- What is EDA?
- An approach to data analysis, largely using graphical techniques to help refine hypotheses and aid in the model building process.
- Some advocates suggest it can be used for hypothesis generation (see the discussion in Bolker Chapter 2 against this).

Objectives of EDA

- Propose/refine models to explain the observed patterns of the data.
- Assess assumptions on which statistical modeling is based
- Provide a context for further data collection.
- Help in determining the appropriate form of statistical modeling (LS, MLE< Bayesian, resampling…).

Check out

http://www.itl.nist.gov/div898/handbook/eda/eda.htm

For more information.

EDA

- EDA emphasizes “robust” and non-parametric approaches to examining the data.

Critiques of EDA (see Bolker chapter 2)

- "Data-Dredging". Using it as "free lunch" with respect to *a posteriori* hypothesis testing.

- Observing patterns that are not real.

  " Under torture, the data readily yields false confessions"

  (MainDonald & Braun 2003)

Suggestions

Ben Bolker suggests an honest approach: prior to examining your data you write down a list of the ecological patterns you are looking for so that you can distinguish between:

1) Patterns you were initially looking for
2) Unanticipated patterns that answer the same question in different ways.
3) Novel (but likely spurious) patterns.

Suggestions

Subset your data:

The other useful approach is to use a random subset of all of your data (No more than 40% of it) for data exploration, and then you can perform the model fitting based on the entire data set (or better yet the remaining 60%).

Of course, this only works if you have enough data to do so!

EDA: numerical summaries vs. graphical summaries

- While point estimates such as means, medians, sd, var, CV are all useful they will not tell you everything.
- Indeed they can lead you astray!!!
- It is imperative to LOOK at your data, at the very least for outliers, if not in an exploratory manner.

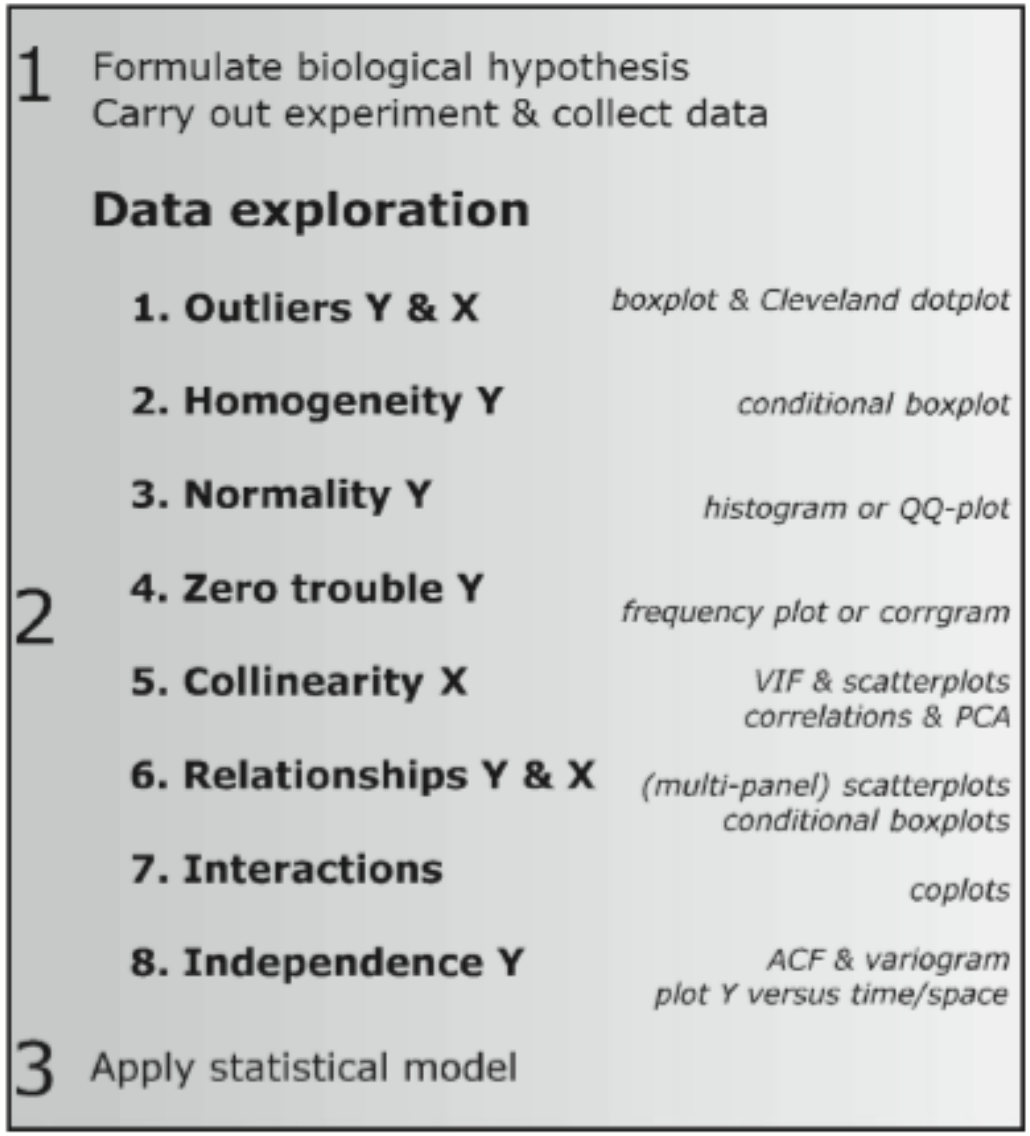


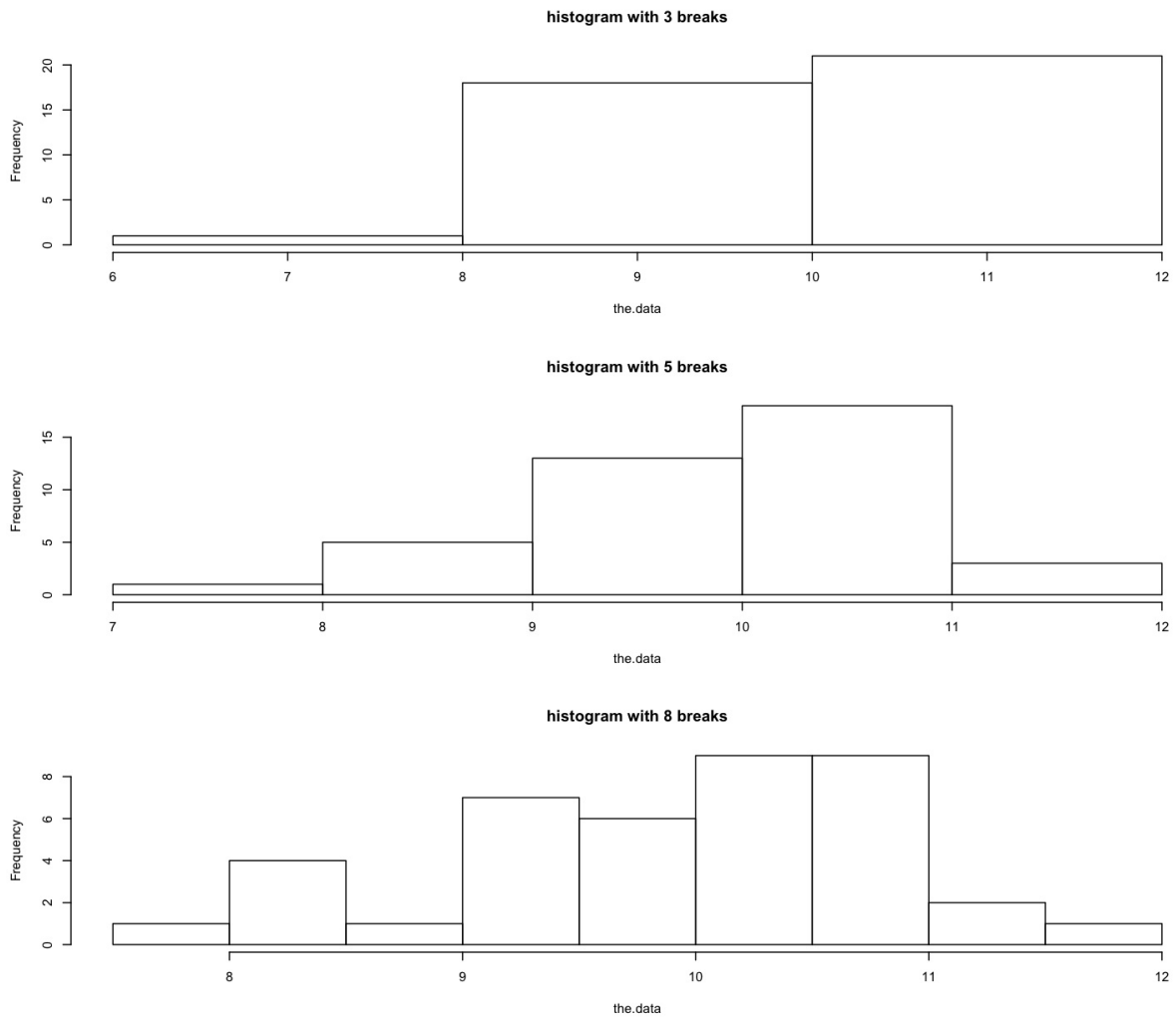
Fig. 1. Protocol for data exploration.

Zuur et al.

Some of the tools: histogram

- The starting point for each variable.
- Extremely useful, BUT be careful about over-interpreting the shape of the data, especially with small sample sizes.

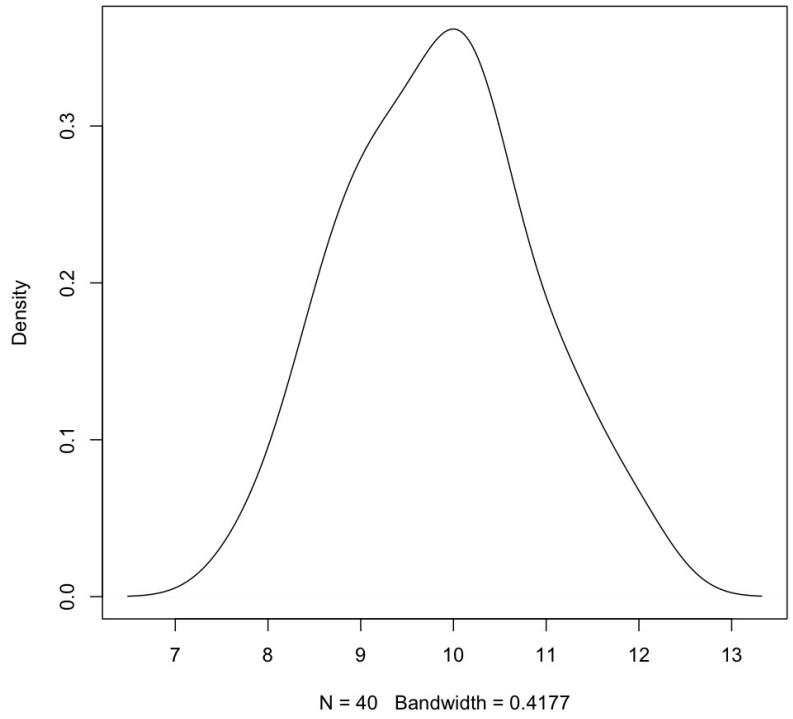
The same data on a histogram



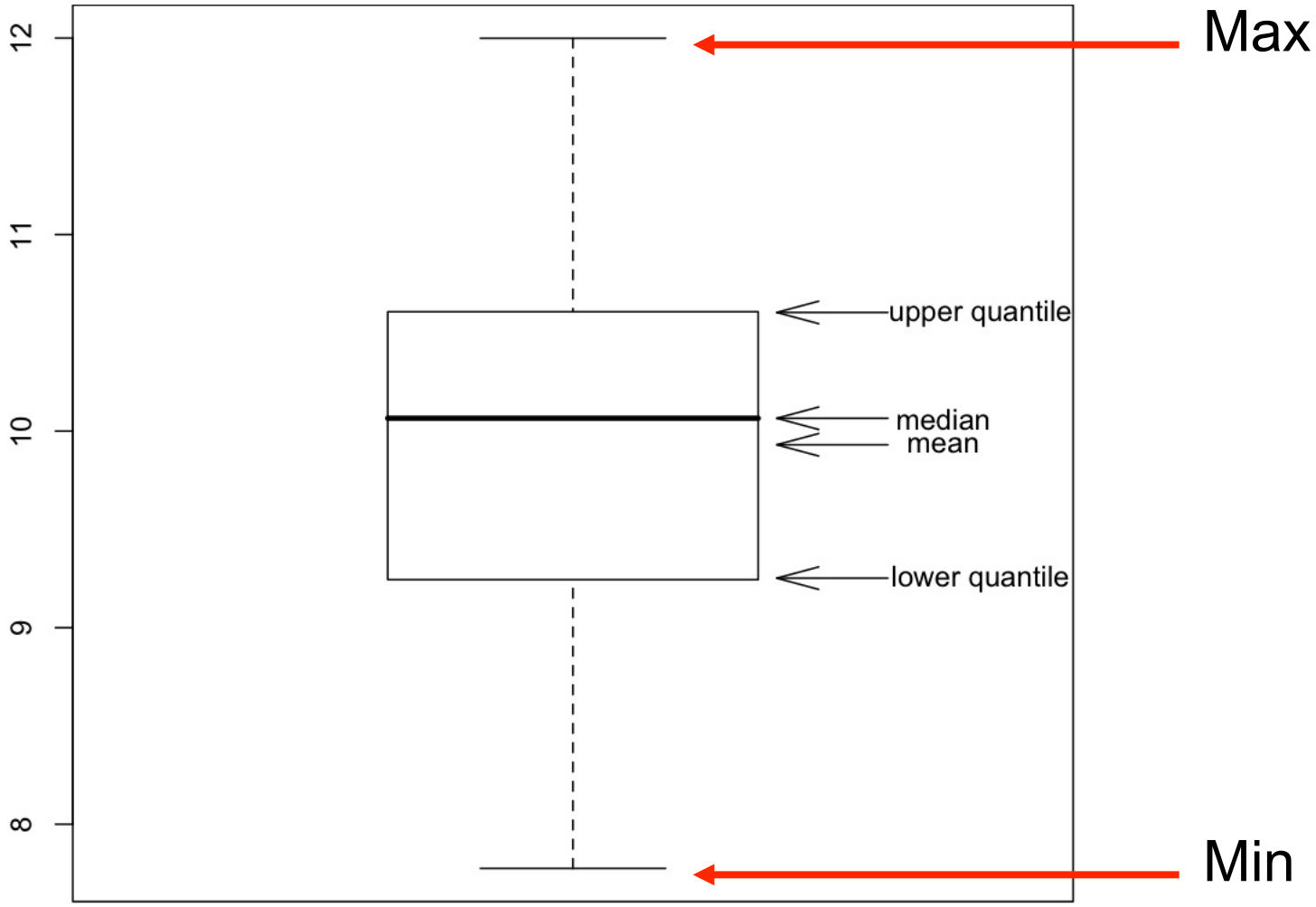
Alternative to histograms

- Histograms are considered a crude form of a density estimate.

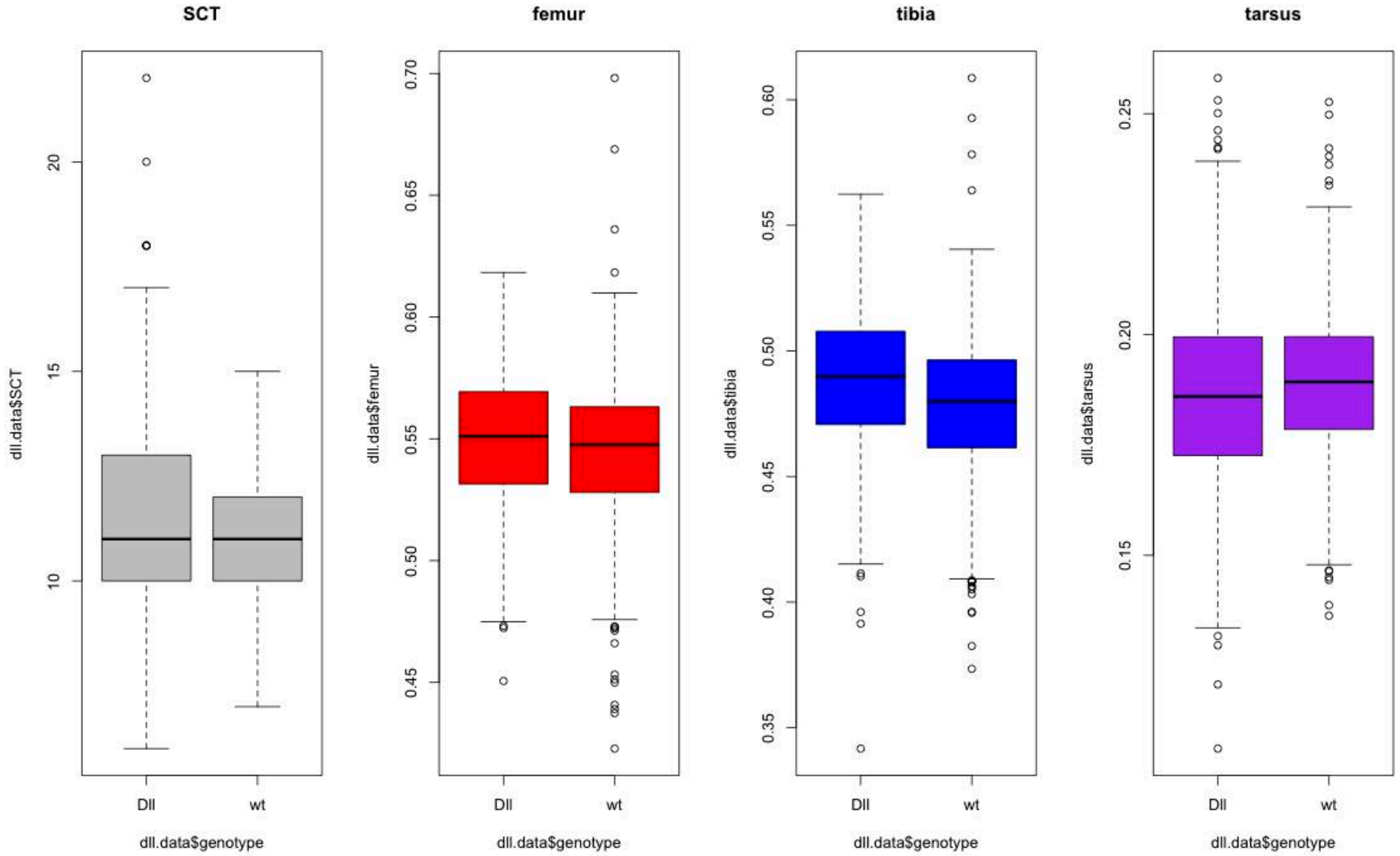
Instead we can actually plot an approximate density for the data



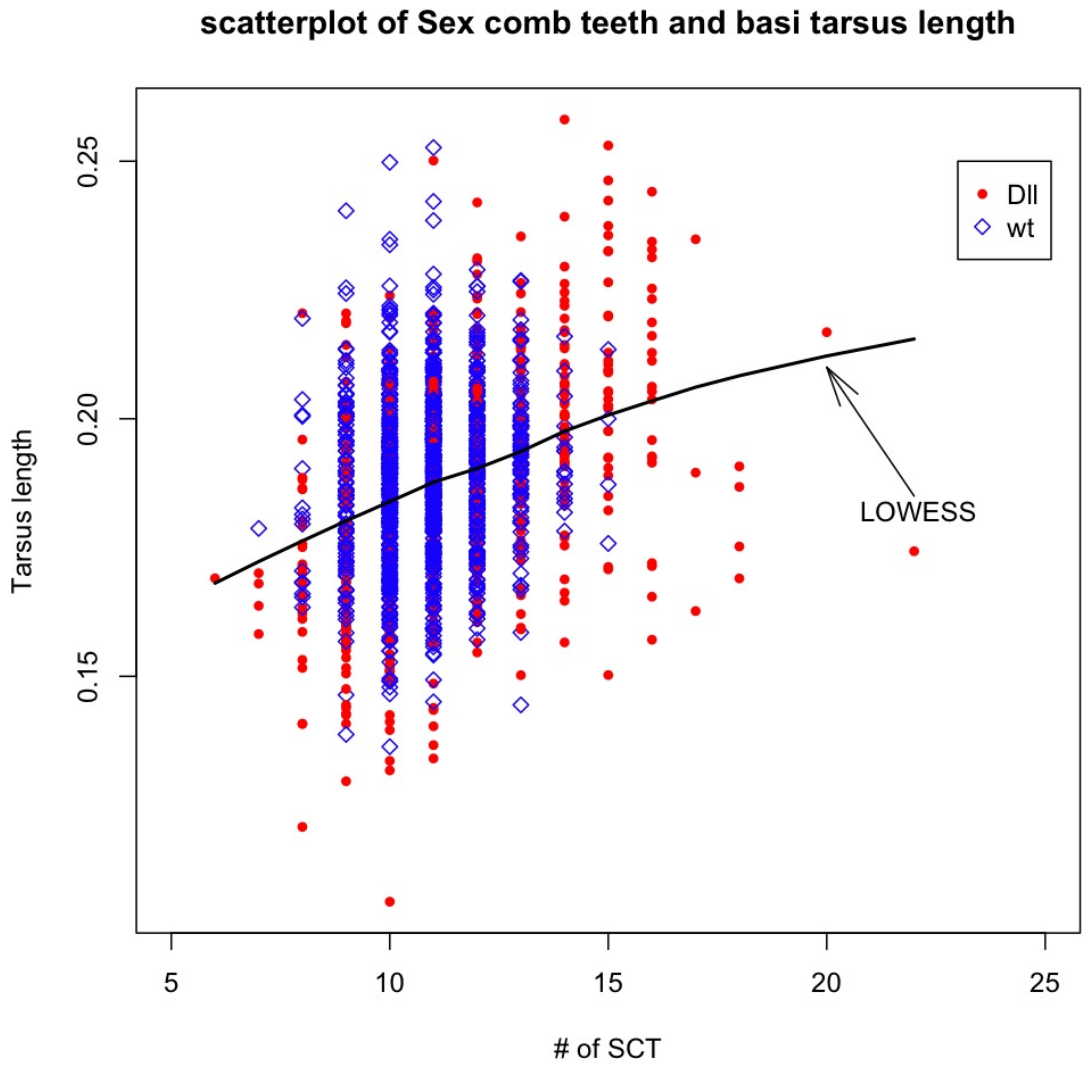
The boxplot


12
11
10
9
8
Max
upper quantile
median
mean
lower quantile
Min

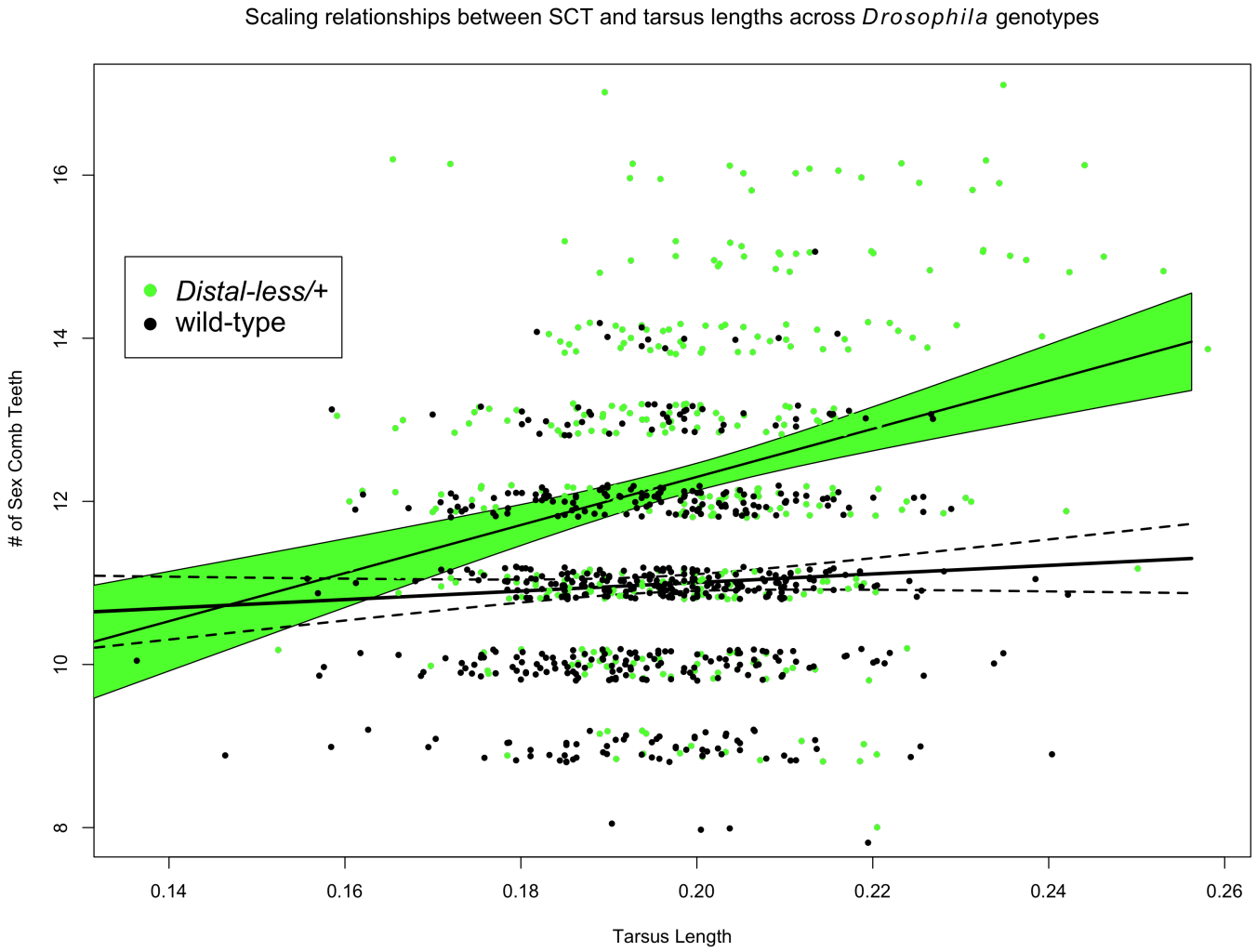
The boxplot


SCT
femur
tibia
tarsus
dll.data$SCT
dll.data$femur
dll.data$tibia
dll.data$tarsus
10
15
20
0.45
0.50
0.55
0.60
0.65
0.70
0.35
0.40
0.45
0.50
0.55
0.60
0.15
0.20
0.25
Dll
wt
dll.data$genotype

Bi-variate scatterplot

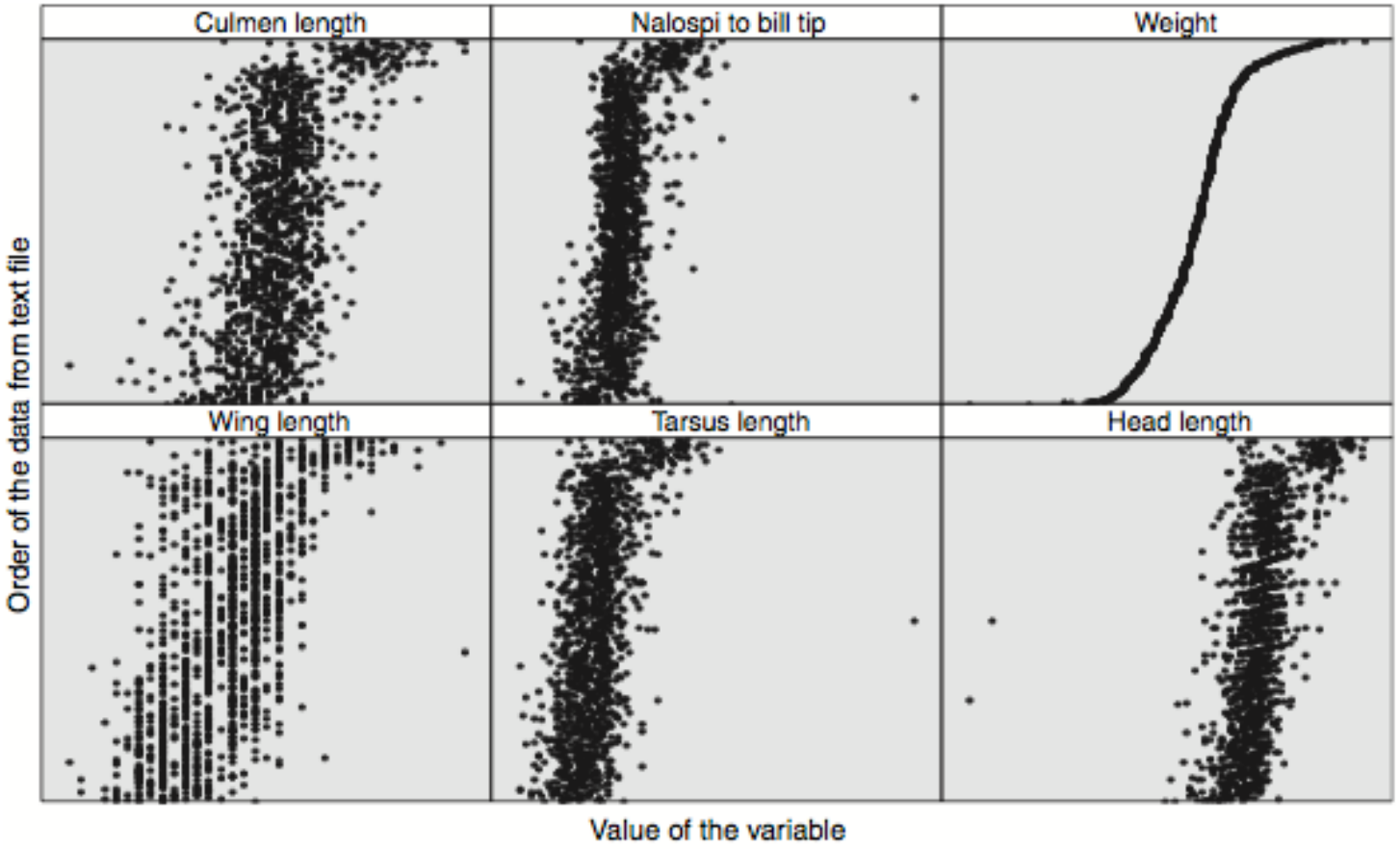


This would not be considered EDA.. Why not?



once you've done linear regression, you've actually fit a model, no longer 'exploratory'

How is this valuable for EDA?



From Zuur et al.

Many other types of plots can be useful depending on your data.

- Chapter 2 of the Bolker book offers many examples.
- Also check out the EDA online lecture for other ideas.

The data we will use for today

EVOLUTION & DEVELOPMENT 7:2, 89–100 (2005)

Evidence for canalization of *Distal-less* function in the leg of *Drosophila melanogaster*

Ian Dworkin[1]

Department of Zoology, University of Toronto, Toronto, ON, Canada M5S 3G5

Correspondence (email: I_Dworkin@ncsu.edu)

[1]Present address: Department of Genetics, North Carolina State University, 3632 Gardner Hall, Raleigh, NC 27695-7614, USA

The purpose

- I wanted to test a model that was proposed to explain the evolution of canalization, or the buffering of traits from new mutations or environmental heterogeneity.

The prediction

- If the hypothesis was correct, we would predict a positive correlation between the response of different genotypes to environmental and genetic perturbation.

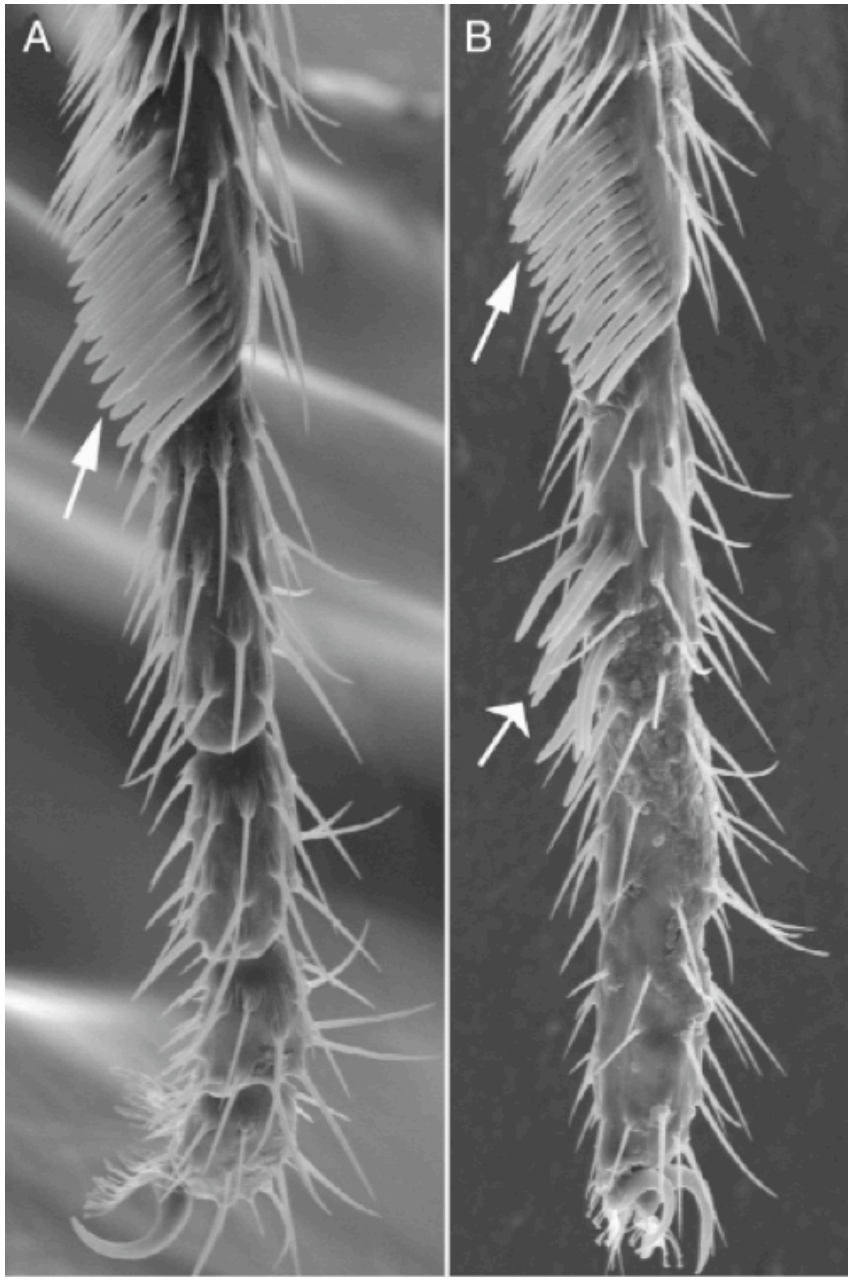
The experiment

- ~ 27 naturally derived strains were sampled from (near Toronto, Ont, Algonquin Park, Ont and world wide samples).
- A single mutation (*Dll*^{B}) was introduced into each of these strains.
- These strains were then reared at two temperature regimes, one “normal” one “stressful”.

Response variables

- I counted the number of sex comb teeth (SCT), a secondary sexual structure used by males to hold onto females during copulation.
- I also measured several length measures on the pro-thoracic leg of the fly (the SCT is located on the basi-tarsus of this leg).

Sex comb teeth



Simpson's Paradox: independent variables (and their responses), when plotted together, could give the appearance of a positive correlation. But when you break down the indep variables into separate groups, there's actually negative correlations within each variable (see Ian's R code in this lecture folder)

Let's start looking at the data, and some R code…..