Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com
Exploring

Exploratory Data Analysis (John W. Tukey, *EDA*)

- Summaries
- Transformations
- Smoothing
- Robustness
- Interactivity

What EDA is not ...

- Letting the data speak for itself
- Fishing expeditions
- Null hypothesis testing

Qualitative Data Analysis

- Mixed methods
- Old wine in new bottles
“Probability modelers seem to want to believe that their models are entirely correct. Data analysts regard their models as a basis from which to measure deviation, as a convenient benchmark in the wilderness, expecting little truth and relying on less.”

Tukey (1979)
Exploring

Summaries

Letter values

M (median) sort and split batch
H (hinges) split each half as if it were a new batch
E (eighths) split again, and so on ...

Medians and hinges yield a 5-number summary

1. lower extreme
2. lower hinge
3. median
4. upper hinge
5. upper extreme

H-spread is (upper hinge – lower hinge)
Range is (upper extreme – lower extreme)
Exploring

Summaries

What letter values reveal

Symmetry

Outliers

A Step is 1.5 times H-spread
Inner fences are 1 step outside hinges
Outer fences are 2 steps outside hinges
Adjacent values are those at each end closest to, but still inside inner fences
Outside values are between inner fence and neighboring outer fence
Far out values are beyond outer fences toward extremes
Exploring

Transformations

Tukey Ladder of Powers (re-expressions)

Assume data are positive, or use $X + 1$ if non-negative

Tukey formula

$X \mapsto X^p$

Box & Cox formula (derived from Tukey’s idea)

$X \mapsto (X^p - 1) / p$

Values of $p$

- $p = 2$ yields $X^2$
- $p = 1$ yields $X$
- $p = .5$ yields $\sqrt{X}$
- $p = 0$ yields $\log(X)$
- $p = -1$ yields $1 / X$

For Box & Cox formula

- $p = 0$ yields $\log(X)$ because $\lim_{p \to 0} (X^p - 1) / p = \log(X)$
- Also, dividing by $p$ in Box & Cox formula preserves polarity of $X$

Ascending the ladder ($p > 1$) spreads out large values and compresses small values.

Descending the ladder ($p < 1$) compresses large values and spreads out small values.
Exploring

Transformations

Dealing with skewness

Positive skew: descend the ladder \((p < 1)\)
Negative skew: ascend the ladder \((p > 1)\)
Exploring Transformations

Wilkinson, Blank, & Gruber (1996)
Exploring

Transformations

Spread-level plot

Divide batch into quintiles and plot H-spread against median

1 – slope of line is estimate of \( p \)

In this case, \( p = 0 \) is best choice
Exploring

Smoothing

See any pattern here?

Day

Temperature

Copyright © 2016 Leland Wilkinson
Exploring

Smoothing

Give yourself a medal if you saw this

Velleman & Hoaglin (1981)
Exploring

Smoothing

Data = smooth + rough

Data = fit + residuals

Fit a model
Compute residuals
Examine residuals for systematic variation
If residuals look nonrandom, fit a model to the residuals
Iterate
Exploring

Robustness

Tukey was skeptical regarding Gaussian assumption
Inspired a search for statistical estimators that are robust against outliers and other forms of contamination
Simple location estimators involved trimming outliers
  Median
  Winsorizing
  Trimmed mean
Others (Tukey, Hampel, ...) involved weighting functions
Peter Huber developed maximum-likelihood-like methods
Exploring

Interactivity

Linking
Brushing
Projecting

Tukey, Friedman, Fisherkeller: Prim9
https://www.youtube.com/watch?v=B7XoW2qiFUA

Tukey and Friedman: Projection pursuit
https://www.youtube.com/watch?v=n5i9RLCe1rQ
Exploring

Qualitative Data Analysis (QDA)

The QDA movement is a reaction against...
- Quantitative analysis (mathematics in general, statistics in particular)
- Scientific objectivism, realism, and positivism
- Peer review (controversial within QDA community)
- Educational testing

Subjectivity
- Hermeneutics
  - Translational
- Postmodernist
  - The researcher constructs own reality that others may not share
- Reliance on “trustworthiness” instead of formal measures of validity
  - Credibility, dependability, auditability, confirmability, corroboration
- Focus on symbolic interpretations of icons (text, videos, …) leads to “mixed methods”

Fluidity
- No predefined measures or hypotheses
- Progressive data collection and coding leads to “grounded theory”

Politics
- Peculiar QDA journals
- Activism in academic departments
Exploring

Qualitative Data Analysis

Nothing new here

- Introspection (Wundt, ...)
- Clinical observation (Freud, Piaget, ...)
- Personal knowledge (Polanyi, ...)
- Participant observation (Malinowski, Mead, ...)
- Community psychology interviewing (Sarason, Levine, Kelly, ...)
- Group dynamics (Lewin, Bales, Slater, ...)

Bottom line:

If you can’t quantify or qualify something, you don’t understand it

- In science, understanding means being able to communicate to a rational person
- In religion, understanding is a non-cognitive experience of the transcendent
- In aesthetics, understanding is a judgment of taste (Kant)
  
But you can’t build a science on subjective or non-cognitive foundations

Quantification doesn’t mean simply assigning a number

- It can mean “these two things are not comparable”
- Or, “this is greater than that”
- Or, “these two things are related”
Exploring

Qualitative Data Analysis Alternatives

Text analysis (Shepard, Rosenberg, ...)

Collect the data through simple comparisons (no numbers)
Scale them by exploiting distance and ordering constraints
Exploring

Qualitative Data Analysis Alternatives

Sequence analysis (Agrawal A priori algorithm)

Association rules
Exploring

Qualitative Data Analysis Alternatives

Network analysis

No numbers here

Exploring

Qualitative Data Analysis Alternatives

Innovative experimental paradigms

No numbers used here

Color vision and hue categorization in young human infants. Bornstein, Marc H.; Kessen, William; Weiskopf, Sally

"Infant looking at shiny object" by Mehregan Javanmard, Wikipedia
Exploring

References


