# Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com

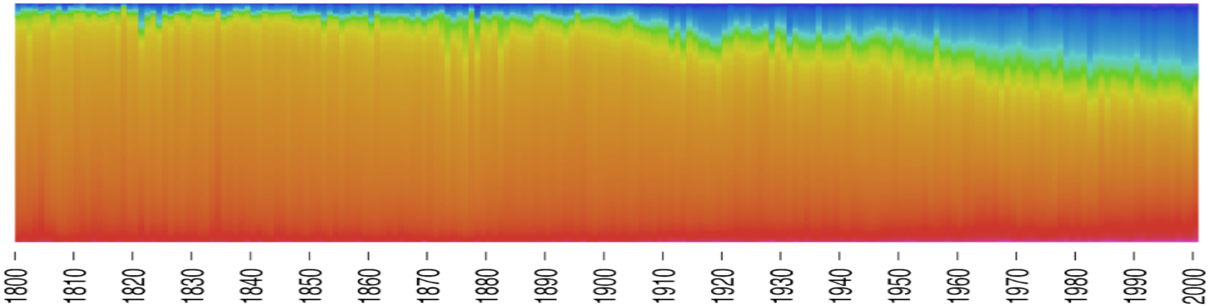
# Data Analysis

What is data analysis?

- Summaries of batches of data
- Methods for discovering patterns in data
- Methods for visualizing data

Benefits

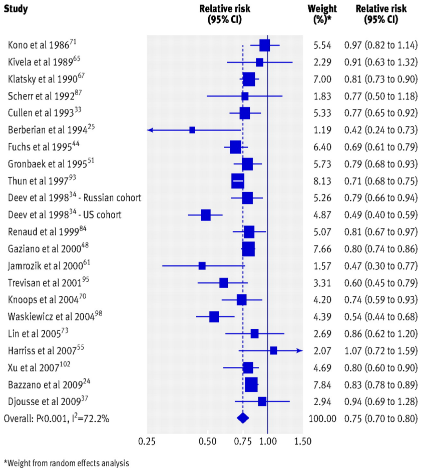
- Data analysis helps us support suppositions
- Data analysis helps us discredit false explanations
- Data analysis helps us generate new ideas to investigate



http://blog.martinbellander.com/post/115411125748/the-colors-of-paintings-blue-is-the-new-orange

# Statistics

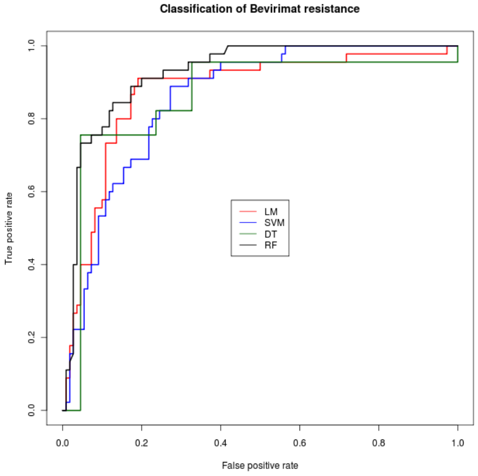
- What is (are) statistics?
  - Summaries of samples from populations
  - Methods for analyzing samples
  - Making inferences based on samples
- Benefits
  - Statistics help us avoid false conclusions when evaluating evidence
    - Statistics protect us from being fooled by randomness
    - Statistics help us find patterns in nonrandom events
    - Statistics quantify risk
    - Statistics counteract ingrained bias in human judgment
  - Statistical models are understandable by humans



http://www.bmj.com/content/342/bmj.d671

# Machine Learning

- What is machine learning?
  - Data mining systems
    - Discover patterns in data
  - Learning systems
    - Adapt models over time
- Benefits
  - ML helps to predict outcomes
    - ML often outperforms traditional statistical prediction methods
  - ML models do not need to be understood by humans
    - Most ML results are unintelligible (the exceptions prove the rule)
    - ML people care about the quality of a prediction, not the meaning of the result
  - ML is hot (Deep Learning!, Big Data!)



http://swift.cmbi.ru.nl/teach/B2/bioinf_24.html

# Course Outline

1. Introduction
2. Data
3. Visualizing
4. Exploring
5. Summarizing
6. Distributions
7. Inference
8. Predicting
9. Smoothing
10. Time Series
11. Comparing
12. Reducing
13. Grouping
14. Learning
15. Anomalies
16. Analyzing

# Data

What is (are) data?

- A datum is a given (as in French donnée)
  - data is plural of datum
- Data may have many different forms
  - Set, Bag, List, Table, etc.
  - Many of these forms are amenable to data analysis
  - None of these forms is suitable for statistical analysis
- Statistics operate on variables, not data
  - A variable is a function mapping data objects to values
  - A random variable is a variable whose values are each associated with a probability *p* ($0 \le p \le 1$)
- Visualizations operate on data or variables

# Visualizing

Visualizations represent data

- Tallies, stem-and-leaf plots, histograms, pie charts, bar charts, …

Statistical visualizations represent variables

- Probability plots, density plots, …

Statistical visualizations aid diagnosis of models

- Does a variable derive from a given distribution?
- Are there outliers and other anomalies?
- Are there trends (or periodicity, etc.) across time?
- Are there relationships between variables?
- Are there clusters of points (cases)?

# 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

# Summarizing

- We summarize to remove irrelevant detail
  - We summarize batches of data in a few numbers
  - We summarize variables through their distributions
  - The best summaries preserve important information
  - All summaries sacrifice information (lossy)
- Summaries
  - Location
    - Popular: mean, median, mode
    - Others: weighted mean, trimmed mean, …
  - Spread
    - Popular: sd, range
    - Others: Interquartile Range, Median Absolute Deviation, …
  - Shape
    - Skewness
    - Kurtosis

# Distributions

- A probability function is a nonnegative function
  - Its area (or *mass*)  is 1
- Distributions are families of probability functions
- Most statistical methods depend on distributions
  - Nonparametric methods are distribution-free
- The Normal (Gaussian) distribution is most popular
  - Other distributions (Binomial, Poisson, …) are often used
  - We use the Normal because of the Central Limit Theorem
    - Variables based on real data are rarely normally distributed
    - But sums or means of random variables tend to be
    - So if we are drawing inferences about means, Normal is usually OK
    - This involves a leap of faith

# Inference

- Inference involves drawing conclusions from evidence
  - In logic, the evidence is a set of premises
  - In data analysis, the evidence is a set of data
  - In statistics, the evidence is a sample from a population
    - A population is assumed to have a distribution
    - The sample is assumed to be random  (Sometimes there are ways around that)
    - The population may be the same size as the sample (not usually a good idea)
- There are two historical approaches to statistical inference
  - Frequentist
  - Bayesian
- There are many widespread abuses of statistical inference
  - We cherry pick our results (scientists, journals, reporters, …)
  - We didn't have a big enough sample to detect a real difference
  - We think a large sample guarantees accuracy (the bigger the better)

# Predicting

Most statistical prediction models take one of two forms

$y = \Sigma_j(\beta_j x_j) + \varepsilon$ (additive function)

$y = f(x_j, \varepsilon)$ (nonlinear function)

The distinction is important

The first form is called an additive model

The second form is called a nonlinear model

Additive models can be curvilinear (if terms are nonlinear)

Nonlinear models cannot be transformed to linear

Examples of linear or linearizable models are

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$

$y = \alpha e^{\beta x + \varepsilon}$

Examples of nonlinear models are

$y = \beta_1 x_1 / \beta_2 x_2 + \varepsilon$

$y = \log \beta_1 x_1 \varepsilon$

# Smoothing

- Sometimes we want to smooth variables or relations
- Tukey phrased this as
  - data = smooth + rough
  - The smoothed version should show patterns not evident in raw data
- Many of these methods are nonparametric
- Some are parametric
- But we use them to discover, not to confirm

# Time Series

Time series statistics involve random processes over time

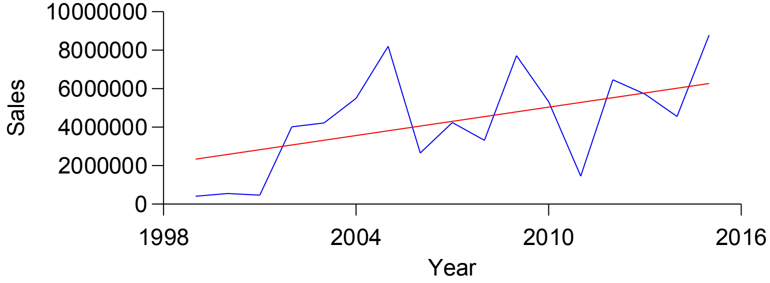
Spatial statistics involve random processes over space

Both involve similar mathematical models

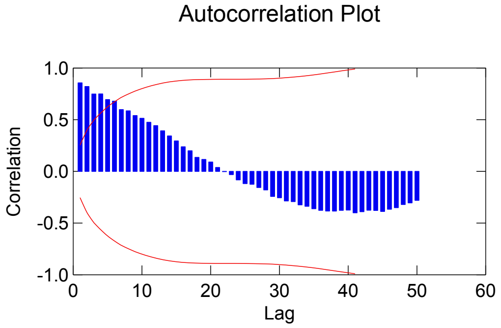
- When there is no temporal or spatial influence, these boil down to ordinary statistical methods

DO NOT USE i.i.d. methods on temporal/spatial data

- These require stochastic models, not "trend lines"
- measurements at each time/space point are not independent



Quarterly US Ecommerce Retail Sales, Seasonally Adjusted



# Comparing

- Statistical methods exist for comparing 2 or more groups
- The classical approach is Analysis of Variance (ANOVA)
  - This method invented by Sir Ronald Fisher
  - It revolutionized industrial/scientific experiments
    - The researcher was able to examine more than one treatment at a time
  - With only two groups, results of Student’s *t*-test and *F*-test are equivalent
- Multivariate Analysis of Variance (MANOVA)
  - This is ANOVA for more than one dependent variable (outcome)
- Hierarchical modeling is for nested data
  - There are several forms of this multilevel modeling

# Reducing

Reducing takes many variables and reduces them to a smaller number of variables

There are many ways to do this

- Principal components (PC) constructs orthogonal weighted composites based on correlations (covariances) among variables
- Multidimensional Scaling (MDS) embeds them in a low-dimensional space based on distances between variables
- Manifold learning projects them onto a low-dimensional nonlinear manifold
- Random projection is like principal components except the weights are random.

# Grouping

- We can create groups of variables or groups of cases
- These methods involve what we call Cluster Analysis
  - Hierarchical methods make trees of nested clusters
  - Non-hierarchical methods group cases into *k* clusters
    - These *k* clusters may be discrete or overlapping
- Two considerations are especially important
  - Distance/Dissimilarity measure
  - Agglomeration or splitting rule
- The collection of clustering methods is huge
  - Early applications were for numerical taxonomy in biology

# Learning

Machine Learning (ML) methods look for patterns that persist across a large collection of data objects

ML learns from new data

Key concepts

- Curse of dimensionality
- Random projections
- Regularization
- Kernels
- Bootstrap aggregation
- Boosting
- Ensembles
- Validation

Methods

- Supervised
  - Classification (Discriminant Analysis, Support Vector Machines, Trees, Set Covers)
  - Prediction (Regression, Trees, Neural Networks)
- Unsupervised
  - Neural Networks
  - Clustering
  - Projections (PC, MDS, Manifold Learning)

# Anomalies

- Anomalies are, literally, lack of a law (*nomos*)
- The best-known anomaly is an outlier
    - This presumes a distribution with tail(s)
    - All outliers are anomalies, but not all anomalies are outliers
    - Identifying outliers is not simple
        - Almost every software system and statistics text gets it wrong
- Other anomalies don’t involve distributions
    - Coding errors in data
    - Misspellings
    - Singular events
- Often anomalies in residuals are more interesting than the estimated values

# Analyzing

- What Statistics is not
  - mathematics
  - machine learning
  - computer science
  - probability theory
- Statistical reasoning is rational
  - Statistics conditions conclusions
  - Statistics factors out randomness
- Wise words
  - David Moore
  - Stephen Stigler
  - TFSI

# References

Statistics

- andrewgelman.com
- statsblogs.com
- jerrydallal.com

Visualization

- flowingdata.com
- eagereyes.org

Machine Learning

- hunch.net
- nlpers.blogspot.com

Math

- quomodocumque.wordpress.com
- terrytao.wordpress.com

# References

Abelson, R.P. (2005). *Statistics as Principled Argument*. Hillsdale, N.J.: L. Erlbaum.

DeVeaux, R.D., Velleman, P., and Bock, D.E. (2013). *Intro Stats* (*4th Ed*.). New York: Pearson.

Freedman, D.A., Pisani, R. and Purves, R,A. (1978). *Statistics*. New York: W.W. Norton.