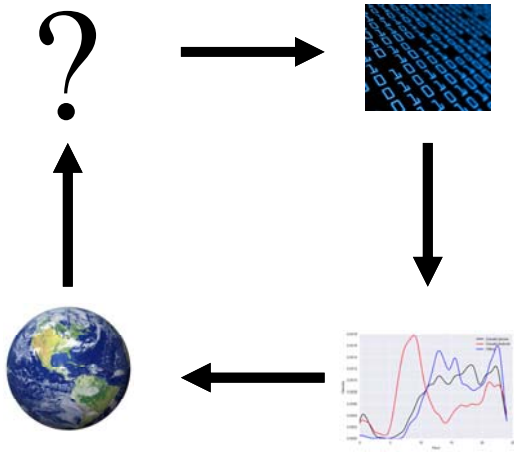
Data Science 100
Lec 7: Prediction, Machine Learning, and Inference

DS100
Spring
2017

Slides by:
Bin Yu
binyu@stat.berkeley.edu



Prediction problems are everywhere

- For satellite images (reflectance measurements at multiple angles and multiple wavelengths), for each pixel, predict the cloud mask
- Using ImageNet (15 mil **human** labelled images with 1000 classes), predict the label with a new image

More prediction examples

- Cancer diagnosis based on genomic measurements

  (BRCA mutation predicts 58% chance of breast cancer for women by age 70)

- Prediction of movie ranking for Netflix users based on past viewing history (for movie recommendation)

- Prediction of fMRI brain signals from input movie clips – key step in movie reconstruction

Presented clip

Clip reconstructed from brain activity

Ultimate importance of prediction

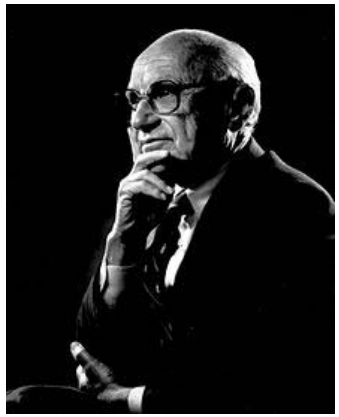
Future holds the unique and possibly the only purpose of all human activities, in business, education, research, and government alike.

Prediction helps us plan for the future, at qualitative and quantitative levels.

Prediction checks reality

"The only relevant test of the validity of a hypothesis is comparison of prediction with experience."

-- Milton Friedman, Economist (1916-2006)



There is always prediction error

“Occurrences in this domain are beyond the reach of exact prediction because of the variety of factors in operation, not because of any lack of order in nature.”

-- Albert Einstein (1879-1955)

Prediction is a basic tool for science

"The only relevant test of the validity of a hypothesis is comparison of prediction with experience." -- Milton Friedman

- Two contexts for prediction evaluation

1. Replication: similar conditions resulting in similar training data and prediction data (representative of the same population)

2. Extrapolation: related but different conditions or related but different training and prediction data (different populations)

A must-ask question in prediction

- Are we in **replication** or **extrapolation** situations?

  Never exactly in one situation: conditions can't be exactly reproduced (sample prep, operator, instrument, …)

- Its answer has to come from understanding of data collection process and domain knowledge – **humans** have to be in the loop and judgment calls are made, better with a panel than by one person

Human is hard to predict

“Human nature is not amenable to prediction based on the trends or tendencies prevailing at the time. It is amenable to startling creativity of the kind practiced by great artists, directors, writers, musicians, actors, who know how to touch a chord in humans everywhere.

-- Maurice Saatchi, Businessman (1946 -)

Gallup poll: a prediction problem about human behaviors

- Gallup poll population: people with phone numbers
- Goal: predict voters' votes on election day

It works under the assumptions that (1) such people have not changed their votes since the time of the poll, (2) that people were telling the truth, (3) people with phones vote similarly as people who do not have phones, and (4) undecided voters vote similarly as decided voters at polling time.

How do we argue for or against the assumptions

- Domain knowledge – humans collectively make a decision
- No assumptions are correct exactly
- Even wrong assumptions could lead to useful predictions – justified by good prediction performance afterwards

All models (assumptions) are wrong, but some are useful

– attributed to George Box, Statistician (1919-2013)

Example for the replication situation

- Same Gallup poll problem
- Predict votes of people WITH phones

In fact, not exactly… it is a “replication” situation only under assumptions (1,2,4).

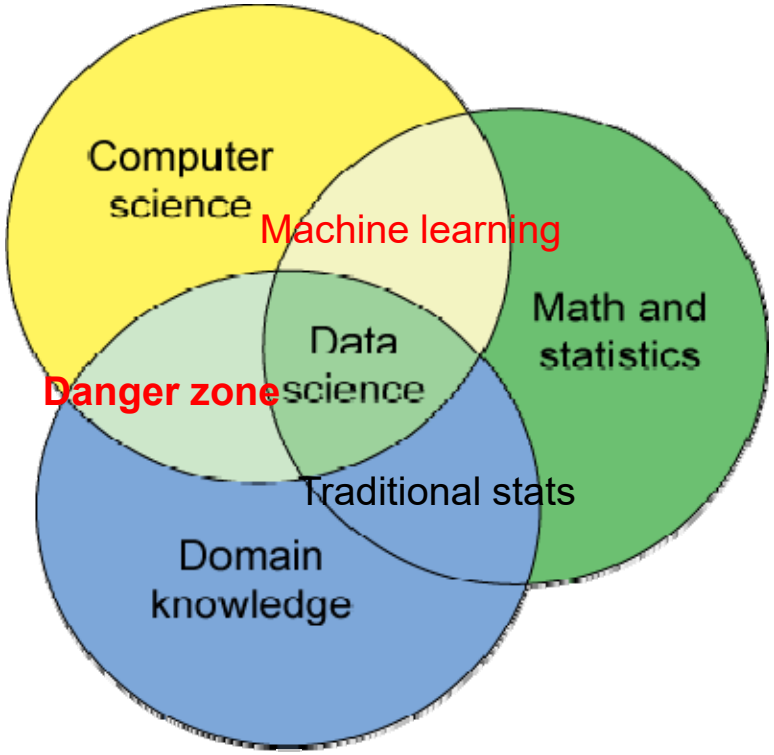
Example for the extrapolation situation

- Same Gallup poll problem
- Predict votes of people WITHOUT phones

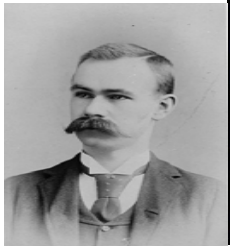
Prediction works under assumptions (1-4)…

Examples of the two situations from class

Recall: what is data science?



Statistician,
Inventor
H. Hollerith



1890's Hollerith Tabulating Machine

Founding father of modern statistics and statistical genetics, R. A. Fisher

Data science is remerging of computational and statistical thinking in the context of domain problems

Machine learning (ML): part of modern statistics and CS

- Traditional statistics has prediction, but not at its center.
- ML: prediction, computation – Cross-validation (CV) workhorse

Prediction + Optimization: cornerstones of ML

- Statistics included prediction, but not at its center

  Cross-validation (CV) was invented by statisticians in the 70's (Stone, 1973; Allen, 1973)

- ML: impressive successes with real-world data problems, especially in IT, with new territories in science and precision medicine and more…

- Computation workhorse: gradient descent and its variants (stochastic gradient descent)…

ML/Stats methods

- Supervised learning (regression and classification):

  (x, y); x-predictor, y – response

  Examples: x: traffic flows in lanes 2 and 3; y: traffic flow in lane 1

  x: pixel values in an image; y: class label (dog, cat, …)

  x: gene expression levels; y: cancer status

  …

  Methods: Simple linear regression, LS, Lasso, Ridge, Nearest Neighbor, Kernel Regression, SVM, Boosting, Neural nets, Deep Learning

ML/Stats methods

- Unsupervised learning (e.g. clustering) just x no response

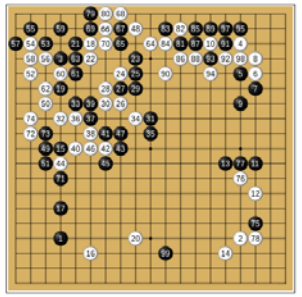
  Examples: x: seal measurements (to find groups)
  x: gene expression levels (to find groups)
  x: pixels on an image (to segment it)

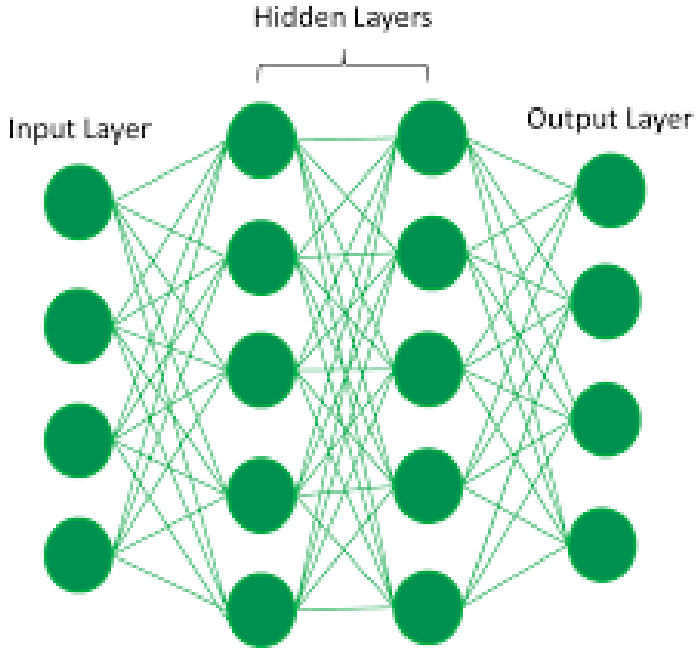
  …

  Methods: K-means, EM, Spectral clustering, Dictionary learning (sparse coding, non-negative matrix factorization

- Graphical models and Bayesian models

Deep learning behind AlphaGo





Pixels

Class labels

An exercise on jargons…

- 1) Will it rain? (Classification)
- 2) How much rain? (Regression)
- 3) Identify similar diseases? (Clustering)
- 4) How do I visualize this high dimensional data? (Dimensionality reduction)
- 5) How do I play the game of Go? (Reinforcement Learning)

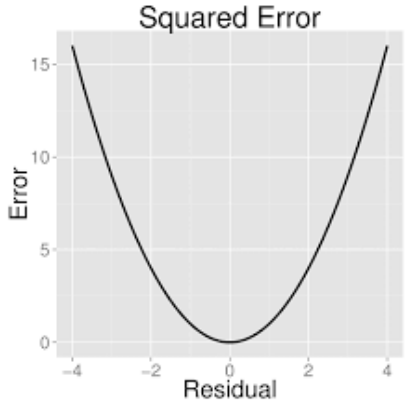
Typical supervised learning set-up

- Data points are treated as exchangeable or homogeneous (each point consisting of one predictor vector x and one response y)

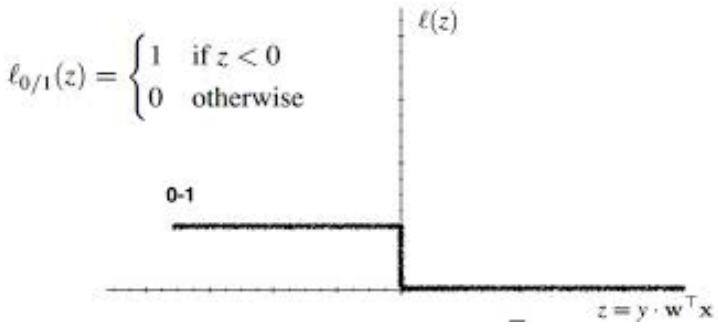
- Goal: build a prediction rule from x to y using data measured by a prediction error (how far the prediction is from the observed response)

- Examples of prediction error for a new pair (x*, y*), use the rule with x* to predict y*

Squared loss



0-1 loss



Reasons for ML success

- Prediction and cross validation are both natural and simple conceptually
- Data availability
- Computing resource availability
- Public domain software

What happened to 2016 election prediction?

- ML/Stats is not magical
- Extrapolation is not the same as replication, and is harder
- One-time shot situation:

  uncertainty needs to be reduced by all means and statistical inference angle is missing…

ML/Stats: frontiers

- Scaling up computation (including communication and memory) to gigantic data
- Dependent/complex data and reinforcement learning (e.g. self-driving car, …)
- **ML becoming more statistical**: Interpretation and statistical inference (e.g. causal inference or AB testing)
- Push back on over-fitting (p-hacking, model-hacking…): all **statistical issues**

Break…

“Inference”, according to Oxford Dictionary

“A conclusion or opinion that is formed because of known facts or evidence”

For data driven opinions/conclusions, known facts or evidence come from data and domain knowledge.

Scientific process: crown jewel of inference

- Glass defined a scientific process as

"the process of determining a quantity Y of interest about some thing X, to a degree of **accuracy** sufficient for **another person to confirm** this quantity Y."

-- from David Glass book "Experimental design for biologist"

Qs: how do we know a prediction is replicated by another lab? They won't get exactly the same number for sure… What is "sufficient accuracy"?

Example: gene mutation and genetic counseling

- One team discovered a gene mutation for a disease (e.g. breast cancer)
- Another team tries to replicate the study
- Will they get exactly the same number? Say 58% of breast cancer rate? (note that 58% is probably a round-off number…)
- When do we declare that the study is replicated or reproduced so a genetic counselor could advise a patient? A decision to make …

Statistical inference

- To assess uncertainty about a conclusion from data

  Or to be more precise,

- To assess uncertainty about quantities drawn from data under a probabilistic data generation mechanism (e.g. sampling from a population, randomization)

How to measure uncertainty?

- Ideally, we (or many labs) **replicate** the data collection process many times and obtain a sampling distribution that describes natural variability

- We can then pull out an uncertainty measure such as a confidence interval from this sampling distribution (Suppose (not real), that the 95% confidence interval from many breast cancer studies is 5%-90% for breast cancer with BRCA positive women – not good enough for a genetics counselor to advise a patient since the breast cancer rate is 12% among women in the general population. 12% is the important subject matter cut-off.)

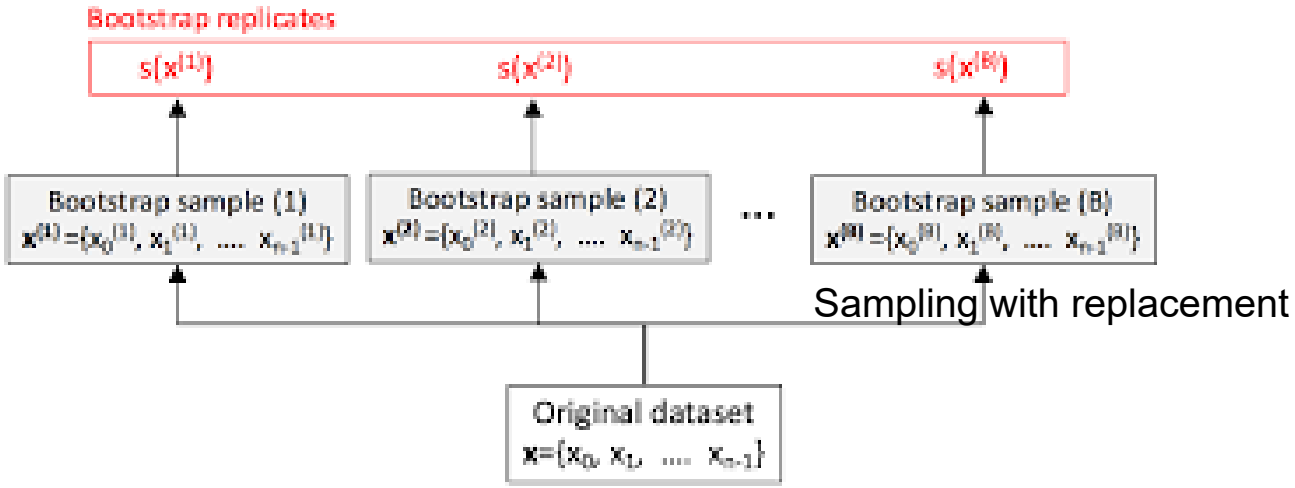
Subject matter matters

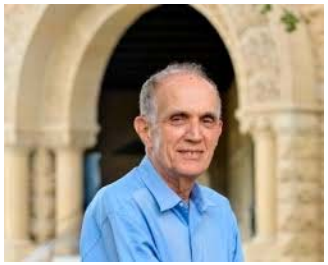
- Why was 5% the right-cut-off? Why not take 10%?

- Statistical inference deals with only the uncontrollable (or random) variability. It does NOT answer how important the data-driven quantity of interest is – subject matter comes in and decides (often through human panels or committees)

In practice, pseudo-replicates are used – Bootstrap replicates

- For “exchangeable” data, randomly draw the same number of data points from the original data WITH replacement to form Bootstrap replicates or samples





Brad Efron (1938-)

Bradley Efron (1979). "Bootstrap Methods: Another Look at the Jackknife". The Annals of Statistics. 7 (1): 1–26.

Case-study: driving decision for J.

- J. drives to work in the morning near Beaumont Ave in Oakland
- J. heard that you had looked at traffic data in DS100 and has come to you for advice:
- J's question: is Lane 4 better to drive in than in Lane 1?

Revisit: traffic data problem

After EDA in Professor Nolan's lecture, let's help J. make a decision (or provide him a confirmatory analysis), while taking into account uncertainty considerations

- Q: is Lane 1 more crowded than Lane 4 in terms of flow during rush hour 7-8 am?

- P: All the work day flows over 60 min intervals (7-8 am) at near Beaumont Ave

- R: what data to use to be "representative"? Under what assumptions?

Data

- Detector: Mainline VDS 400302 - Beaumont Ave
- Data are counts of cars that pass over the detector 7am-8am every work day
- Excludes weekends and holidays

Thanks to Andrew Do

Do we have “exchangeable data”?

- Class discussion
- In practice, a good discussion among team members is necessary
- If you are a one-person team, play many sides yourself – be critical of your own thoughts

Why do we need exchangeability?

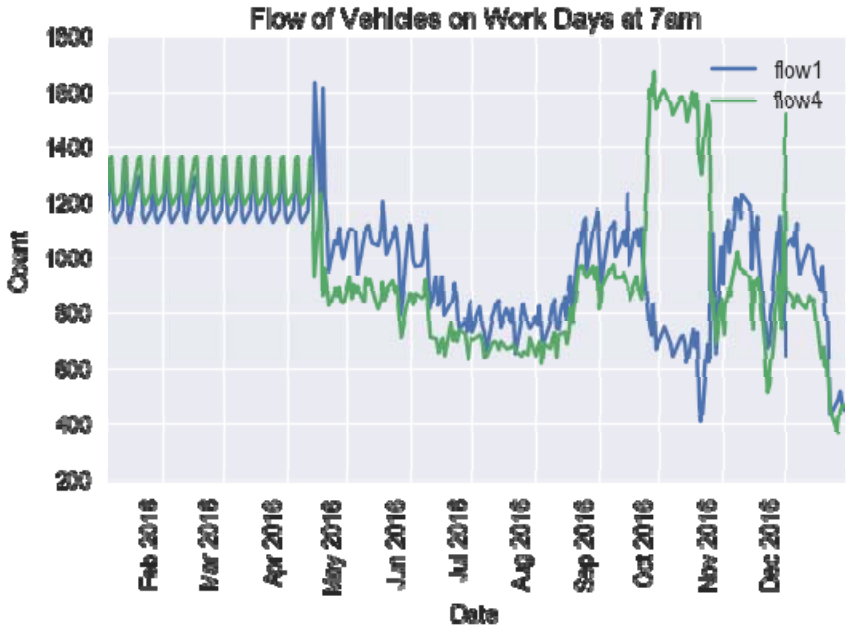
- It ensures as much as possible that the bootstrap samples are “replications” of the original data while using the original data
- It ensures that drawing without replacement of each data point is legit
- For time series data, people do block-bootstrap samples to keep (to a certain extent) the time-structure in the original time series

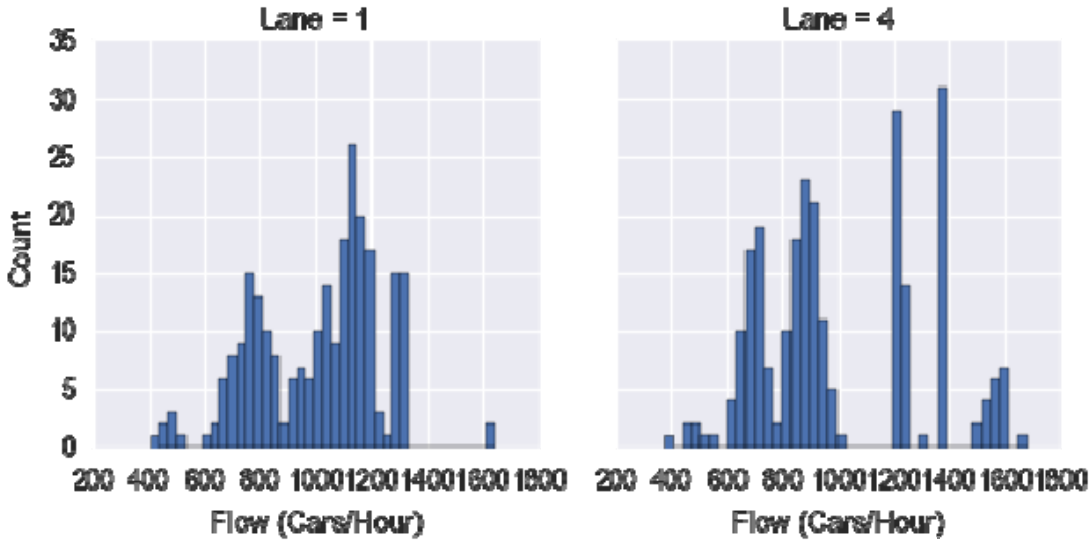
Translating J's question into a statistical Q

- What does he mean by "better"?
- Less crowded on average?
- More controlled arrival time at work? Less surprises with his travel time?

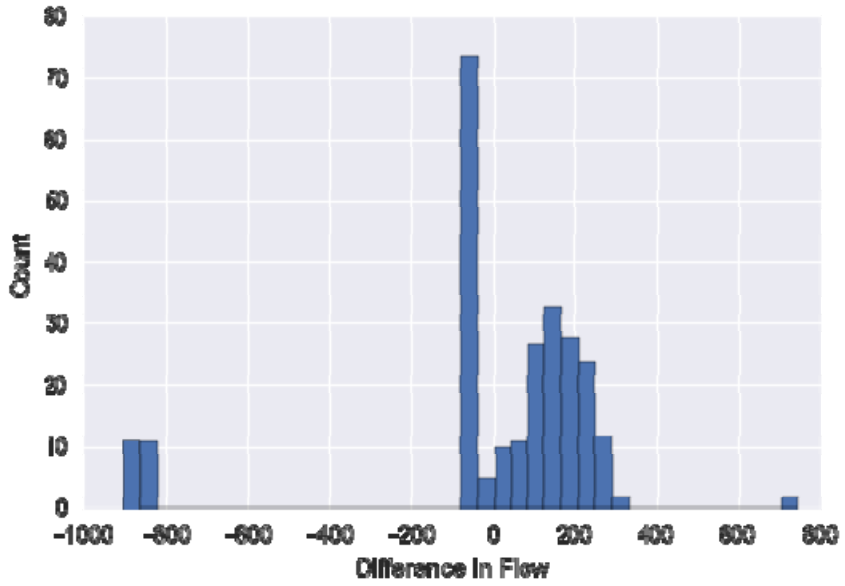
Statistical Q:
is the mean difference larger than zero?

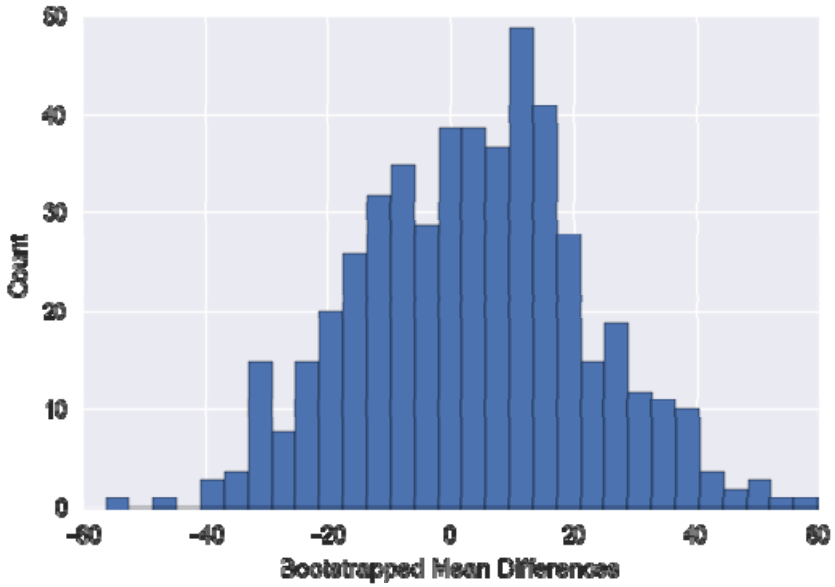
Visualizing data





Why bootstrap the differences?





Drawing conclusion and communicating to J.

- From our bootstrap procedure, we get a one-sided 95% confidence interval for the mean difference: $(-29.31, \infty)$, which includes 0.
- So the Null hypothesis of no mean difference accepted at 5% -- not enough evidence that the mean diff is not zero. It does not prove it is zero, however.
- Tell J: no difference between Lane 1 and Lane 4?

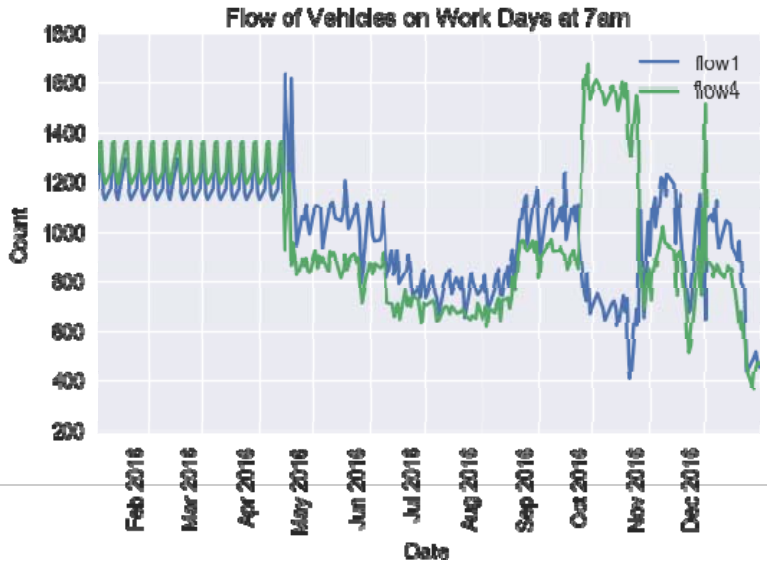
What would you say?

Drawing conclusion and communicating to J.

- From our bootstrap procedure, we get a one-sided 95% confidence interval of $(-29.31, \infty)$

- Tell J: no mean difference between Lane 1 and Lane 4?

What would you say?



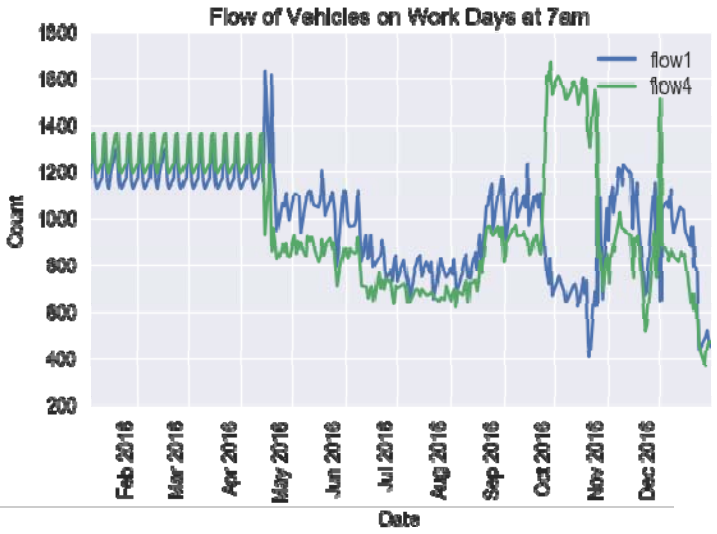
Drawing conclusion and communicating to J.

- From our bootstrap procedure, we get a one-sided 95% confidence interval of $(-29.31, \infty)$

- Tell J: no mean difference between Lane 1 and Lane 4?

What would you say?

Confidence interval is not the appropriate tool for J. to make a decision.



Take-home message

- Confidence interval is a good formal summary of **consistent** data analytics based on domain knowledge and visualization

- Not otherwise; it can be misleading without other data evidence including visualization and domain knowledge

Concept/term summary

- Prediction (prediction error, replication vs extrapolation)
- Statistical machine learning (supervised (regression, classification), unsupervised (clustering), reinforcement learning)
- Statistical inference (uncertainty, exchangeable data, bootstrap)
- Translation of context Q to stat Q – not unique
- Decision making: inference only part of the evidence, not all of it (other data evidence includes visualization and domain knowledge)

Extra slides for your reference

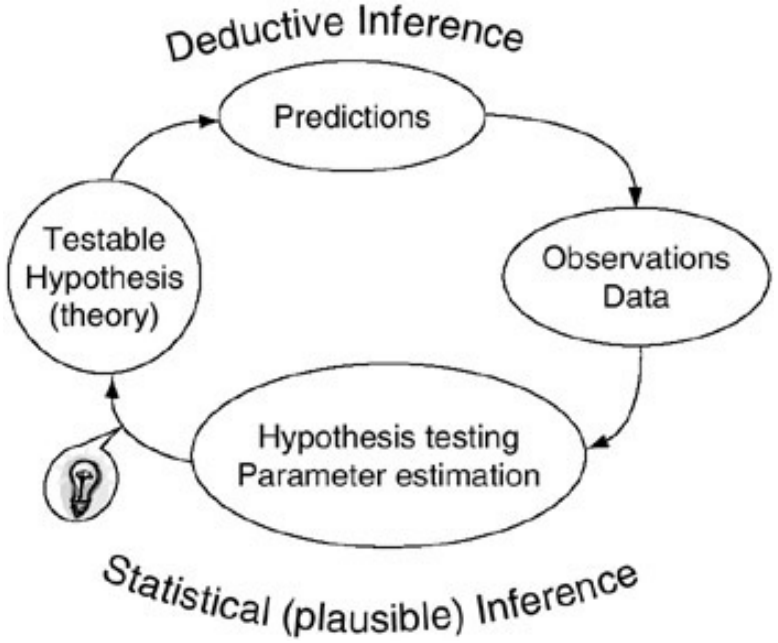
Statistical Q: is the mean difference larger than zero?

```
>occ.diff=traffic[,2]-traffic[,3]
> hist(occ.diff)
> mean(occ.diff)
[1] 3.241567
>m500=seq(1:500)
>for (i in 1:500){sam=sample(occ.diff,length(occ.diff),replace=T);
m500[i]=mean(sam)}
> hist(m500)
>m500.sort=sort(m500)
> m500.sort[25]
[1] 2.731151
```

So one-sided 95% percent confidence interval is (2.731151, \infty)

Accuracy: when is it good enough?

Bench mark: reference distribution driven by a random mechanism

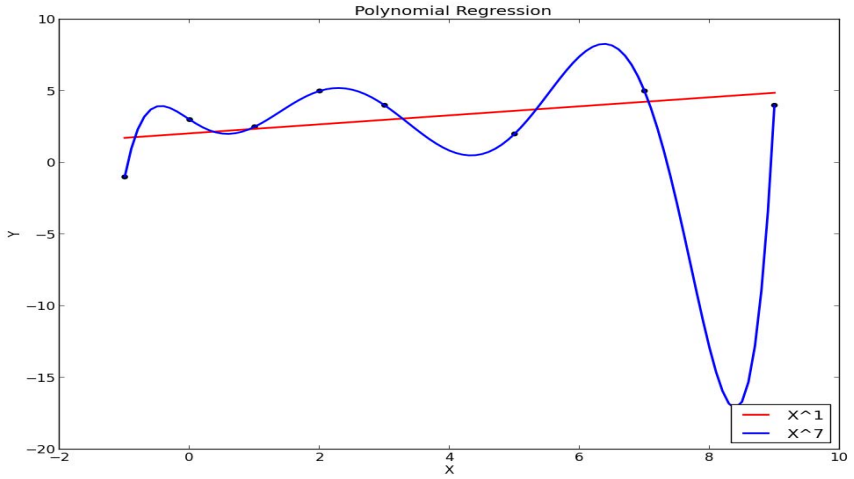


Choosing a prediction rule from a class

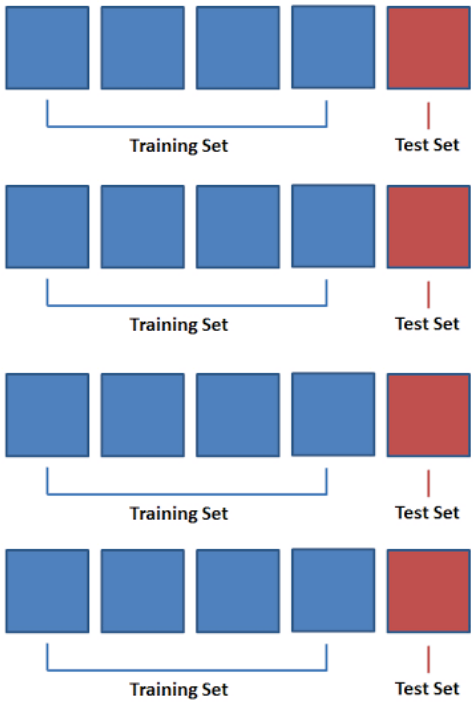
- One often considers many prediction rules indexed by a tuning parameter that controls the complexity (e.g. degree of a polynomial)

- Intuition: the more complex the rules, the better the fit to training data, but not necessarily the better the prediction performance – it is more likely to overfit

  - simple
  - complex



Ideally, we replicate the data collection process k times (by k labs, say) and use prediction error to choose the tuning parameter



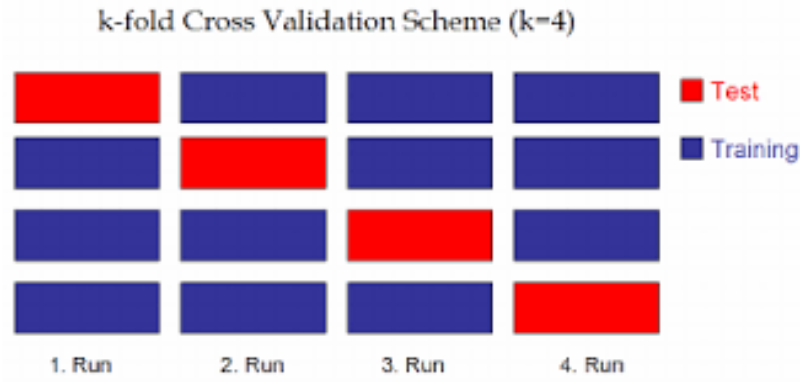
Example: k cancer diagnosis studies by k medical centers

Prediction error

= average of prediction errors from the k labs

k-fold Cross-validation: mimicking “replication” within a data set

- CV idea: create k prediction problems within a data set



CV prediction error = average prediction error over the k runs

- Pseudo-replication of k prediction problems: they should be similar to each other and to the original problem – “representativeness”
- CV prediction error can be bad estimate of replication prediction error because the repeated use of the same data sub-blocks

How good is CV?

- Cross-validation works well for finding the right amount of complexity or choosing the tuning parameter based on data, but the CV prediction error is not always a good estimate of the prediction error (due to often **positive dependence** between the prediction errors over sub-blocks)

- True story: CV gives 90% CV-accuracy with randomly assigned cancer labels with gene expressions as predictors

- Go for proper test prediction error if you can