Small Statistics Big Data Curriculum

C.Foster
05 Oct 2018
Historic

- Database access controlled and complex (Instance / Schema / Table)
  Software Engineer to organize and query data.

- Analysis in proprietary code (FEA/CFD Solvers)
  Analysis Engineer to setup and run solvers

- Statistical Analysis in special programs (JMP/Minitab)
  Statistician run analysis

- Publish results (HTML/Java)
  Programmer to create report or dashboard

No common process or link
Process & Tools
• Get Data
• Analyze and model
• Publish results

Target Audience: Engineers / Technicians / Scientists who are not primarily digital.

Long term control & process stability

Existing processes
- Define
  - Search

- Measure
  - Stitch

- Analyze
  - Science
  - Show

- Improve
- Control
  - Share

Digital Tools

Training Day
1 2 3-4 5
Query → Norm / Imputing → Aggregate

First N - Sampling
- Data Engineering Process
  - Z-Score
  - 0 Replacement

Stratified Sampling
- Applied Statistics Process
  - Min/Max
  - Robust Scaling
  - Imputation

Charts → Tests → Models

Charts
- Scatter / Bar / Line / Combined

Tests
- Data Visualization Process
  - Color by... / Line by... / Shape by...

Models
- Applied Statistics Process
  - Linear and Logistic Regression
  - Classification and Regression Trees

- Z-t Test / $\chi^2$ / FDR

- Median
- IQR

- Mean
Integrate the Statistical / Data Science methods and the Tools and Process Activities
Define the goal in terms of a measurable outcome

**Project Charter**

<table>
<thead>
<tr>
<th>Problem Statement</th>
<th>Goal Statement</th>
<th>VOC:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Team:</td>
<td>Project Information:</td>
<td>Key Metrics:</td>
</tr>
<tr>
<td>Leader:</td>
<td>Project start:</td>
<td>:</td>
</tr>
<tr>
<td>Team member1:</td>
<td>Project end:</td>
<td>:</td>
</tr>
<tr>
<td>Team member2:</td>
<td>Project approach:</td>
<td>:</td>
</tr>
<tr>
<td>Team member3:</td>
<td>Project scope:</td>
<td>:</td>
</tr>
</tbody>
</table>

**Milestones:**

**Signatures:**

---

1. \(N \neq \text{All}\)
2. \(N\) is big
3. Act
4. Feedback

**Four Examples**

- Something is missing
- Statistical assumptions breakdown.
- What is wrong with the measurements? Are we solving the right problem?
- How will the system react and how will it effect the analytic?

Models / Ownership
N≠All
N is big

R = .1

100 points p = 0.936

Not Significant

10000pts p < 2e-16

VERY Significant
Your task is to turn over as few cards as possible to verify whether the following statement is true.

*Every card with a vowel on one side has an even number on the other.*

Which order would you turn over the cards?

1. AB23    4. A2B3
2. AB32    5. A3B2
3. A23B    6. A32B
Your task is to turn over as few cards as possible to verify whether the following statement is true.

*Every part with a crack has corrosion as a cause*

Which order would you turn over the cards?

Which order would you turn over the cards?

**Could a cause be crack?**

*i.e. progressive damage.*
Chasing Noise...

Students Have a Poor Landing

Engine has high fuel consumption one month

Yell at student

...Tell Operator

Next Landing Improves

Fuel Economy Improves

All data are noisy
Interactions with the System

If customer is charged more when they are here
They try to run here
Even if it does not change risk / cost

X is a surrogate for Analytic based on correlations
Customer tunes X and changes correlations to Y

Control is critical
Not predicting weather
Methods and Big Data

- Small ANN
- Medium ANN
- Logistic / Random Forest / Support Vector Machine

Performance vs. Data

Likelihood
Gaussian Process
Random Forests

General use model

Pretty well most of the time

Fails obviously

Includes Uncertainty

Bayesian Calibration of Computer Models

Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife
http://jmlr.org/papers/volume15/wager14a/wager14a.pdf
The Analyst:

Owns the proposal

Owns knowing the data - and quality

Gets the data from the lake

Stitches the sources

Creates the visualization
Training Process

Integrate objectives

Graphical Program Concept

Statistical Concept

Learning questions

Follow an available text
https://www.openintro.org/stat/textbook.php
P-Value Sample

A poll by the National Sleep Foundation found that students on average sleep 7 hours per night.

A sample of 169 students sleep for one night had an average of 6.88 hours and a standard deviation of 0.94 hours.

Assuming that this is a representative random sample, is there sufficient evidence to reject the null hypothesis that students on average sleep 7 hours per night?

What is the p-value for this hypothesis test?

What caveats would you add?

https://www.openintro.org/stat/textbook.php
Student Case Study – West Nile Virus
(with parallels to fleet management)

Setup project requirements and deliverables

Find and manipulate data

Visualize and build models

Validate and Publish Dashboard
Complete process with tools

- Engineers and users are capable of entire analytic process
- Basic statistics knowledge and capability
- Integrated with current improvement process