Why analyze data? How variety in the objectives of analysis points to complementary roles for statistics and data science.

Dan J. Spitzner

Department of Statistics
University of Virginia

October 18, 2017
About the presentation

**Organization:** Short- to medium-length vignettes of varying scope and topics

**What to look for:**

- A thread of applications in *forensic pattern matching*
- The wide variety of *motivations* and *objectives* of data analysis
- Philosophical *criticisms* and *arguments* related to meaning in data analysis
A marketing company has compiled data on a subset of credit-card account-holders, which is to be used to develop a scoring formula with which to target individuals for a promotion.

<table>
<thead>
<tr>
<th>resp</th>
<th>ID</th>
<th>balance</th>
<th>income</th>
<th>age</th>
<th>region</th>
<th>rating1</th>
<th>rating2</th>
<th>rating3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>144</td>
<td>1446</td>
<td>B</td>
<td>69</td>
<td>D</td>
<td>57</td>
<td>22</td>
<td>43</td>
</tr>
<tr>
<td>0</td>
<td>148</td>
<td>23832</td>
<td>C</td>
<td>38</td>
<td>C</td>
<td>74</td>
<td>75</td>
<td>60</td>
</tr>
<tr>
<td>0</td>
<td>149</td>
<td>2407</td>
<td>B</td>
<td>61</td>
<td>D</td>
<td>47</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>152</td>
<td>57983</td>
<td>D</td>
<td>57</td>
<td>F</td>
<td>33</td>
<td>92</td>
<td>96</td>
</tr>
<tr>
<td>0</td>
<td>155</td>
<td>109</td>
<td>A</td>
<td>72</td>
<td>A</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>714</td>
<td>204</td>
<td>A</td>
<td>23</td>
<td>A</td>
<td>42</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>0</td>
<td>715</td>
<td>22847</td>
<td>C</td>
<td>72</td>
<td>D</td>
<td>95</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>1</td>
<td>719</td>
<td>24497</td>
<td>D</td>
<td>69</td>
<td>B</td>
<td>81</td>
<td>72</td>
<td>85</td>
</tr>
<tr>
<td>0</td>
<td>720</td>
<td>142</td>
<td>A</td>
<td>55</td>
<td>B</td>
<td>75</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>0</td>
<td>723</td>
<td>39358</td>
<td>D</td>
<td>57</td>
<td>C</td>
<td>80</td>
<td>96</td>
<td>98</td>
</tr>
</tbody>
</table>

Why analyze data?: Data analysis can be part of a company’s resource investment strategy
Stopping rules in classical hypothesis testing

Collect $x_1, \ldots, x_n$, each $x_i \sim N(\mu, \sigma)$. To test $H_0 : \mu = \mu_0$, use

$$z_{obs}(n) = \frac{\bar{x}_{obs} - \mu_0}{\sigma / \sqrt{n}}$$

**Stopping rule 1:** $\alpha = 0.05$
- Stop collecting observations at $n = 100$
- Reject $H_0$ if $|z_{obs}(100)| > 1.96$

**Stopping rule 2:** $\alpha = 0.05$
- Collect $n = 100$ observations
  - If $|z_{obs}(100)| > 2.18$, stop and reject $H_0$
  - Otherwise, collect another 100 observations
    - If $|z_{obs}(200)| > 2.18$, stop and reject $H_0$

**Conundrum:** If $|z_{obs}(100)| = 2$, significance depends on the experimenter’s thoughts about the future
Albert-László Barabási examined the anonymous logs of millions of mobile phone calls for about four months

- When **people with many links within their community** are removed, the social network does not fail.
- The loss of **people having links outside the immediate community** risks social network disintegration.
- This pattern seems only detectable when examined at a large scale

**Why analyze data?:** Visualization of complex phenomena can generate hypotheses and inspire explanatory investigations
Trigonometric regression

\[ f_j(t) = \begin{cases} \cosine \text{ with period } j/2 \text{ units} & \text{if } j \text{ is even} \\ \sin \text{e with period } (j + 1)/2 \text{ units} & \text{if } j \text{ is odd.} \end{cases} \]
Trigonometric regression

Average temperature across the year.

\[ y_t = \beta_0 + \beta_1 f_1(t) + \beta_1 f_2(t) + \cdots + \beta_k f_k(t) + \epsilon_t \]
Curve decomposition

<table>
<thead>
<tr>
<th>Location</th>
<th>Cumulative</th>
<th>Add</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.81%</td>
<td></td>
<td>19.19%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tilt</th>
<th>Cumulative</th>
<th>Add</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>94.14%</td>
<td></td>
<td>5.86%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Freq. 1</th>
<th>Cumulative</th>
<th>Add</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.29%</td>
<td></td>
<td>0.71%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Freq. 2</th>
<th>Cumulative</th>
<th>Add</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.80%</td>
<td></td>
<td>0.20%</td>
</tr>
</tbody>
</table>
Curve decomposition

All Curves
Var: 293.7

PC1
53.27%

PC2
29.38%

PC3
7.28%
High-dimensional modeling and testing

To analyze a random sample of functions,

\[ X_1(t), \ldots, X_n(t) \] with common domain \( t \in D \),

follow these steps:

**STEP 1:** Apply a **decorrelating** decomposition

| random function \( X_i(t), t \in D \) | high-dimensional vector \( X_i = (X_{i1}, \ldots, X_{ip}), \) indep. \( X_{ij} \) |

**STEP 2:** **Downweight** less interpretable* elements

\[
\|Z\|_w^2 = \sum_{j=1}^{p} w_j \left( \frac{\bar{X}_j - \mu_{0j}}{\sigma_j/\sqrt{n}} \right)^2
\]

*such as a “smoothness” interpretation: under “Sobolev” smoothness, set \( w_j = j^{-1/2} \)
Why analyze data?: Formal inference methods aim to summarize and weigh evidence of some condition.
Bayesian inference

**STEP 1:** Define the phenomenon

**STEP 2:** Express what is already known about the phenomenon probabilistically

\[ \Rightarrow \text{prior probability, } \pi(\theta) \]

**STEP 3:** Express how data are generated probabilistically

\[ \Rightarrow \text{likelihood function, } \pi(Y|\theta) \]

**STEP 4:** Collect the data

**STEP 5:** Update what is known using Bayes’s theorem

\[ \Rightarrow \text{posterior probability, } \pi(\theta|Y) \]

End result is a probabilistic expression of what we know
DeFinetti representations

$X_n = (X_1, \ldots, X_n)$ is a dependent bit sequence

I think I can learn about $X_{n+1}$ from $X_n$

**DeFinetti:**
- There is a parameter $\theta$, defined as $\theta = \lim_{n \to \infty} \bar{X}_n$
- There is a probability distribution $Q(\theta)$ associated with $\theta$
- Conditionally, $X_n|\theta$ is an independent bit sequence
- $P[X_{n+1}|X_n]$ is obtained from the distribution $Q(\theta|X_n)$ given by Bayes’s Theorem

This solves the **induction problem** by connecting past and future through a prior distribution, $Q(\theta)$

A DeFinetti representation is a coherent model for learning
Bullet land matching

A gun barrel’s rifling leaves a unique mark on bullets

**Idea:** Construct a *metric* for comparing striae in bullet lands*

*A “land” is a impression made by the raised portion between groves in a barrel’s rifling*
Bullet land matching

Hare, Hofmann, and Carriquiry’s metric

STEP 1: Crop “shoulders”
STEP 2: Apply smoothing
STEP 3: Collect residuals
Bullet land matching

**STEP 4:** Align residual profiles by minimizing cross-correlation

**STEP 5:** Locate peaks and valleys

**STEP 6:** Find matching striations *via* overlapping intervals
Bullet land matching

Some **features** of aligned profiles

- Maximum consecutive matching striae (CMS)
- Maximum consecutive non-matching striae (CNMS)
- Number of matching striae
- Number of non-matching striae
- Cross-correlation value
- Average squared difference between profiles
- Total heights and depths of matched peaks and valleys
Evaluation: On a test data set . . .

- Every feature performs well individually in distinguishing matches from non-matches
- A decision tree built on the features performs well in distinguishing matches from non-matches
- A random forest performs well in distinguishing matches from non-matches

“. . . we can successfully employ machine learning methods to distinguish matches from non-matches”

–Hare, Hofmann, and Carriquiry
Bullet land matching

**Why analyze data?** to “... eliminate the need for a visual inspection during the matching process and replace it with an automatic algorithm”

**Note:** The objective is **not** to summarize and weigh evidence of some condition:

“Determining a threshold such that [feature] values above the threshold indicate a match with high reliability is beyond the scope of this work, even though it is critically important in practice.”
BP’s oil refinery monitoring

At a BP oil refinery in Washington state, wireless sensors continually monitor the state of the oil-refining process.

- Data from individual monitors may become inaccurate due to the effects of heat and other stresses on the sensors, but the huge number of sensors is able to make up for it.
- By monitoring pipes in this way, BP came to realize that some types of crude oil are more corrosive to its equipment than others.

Why analyze data?: Data streams from a sophisticated monitoring apparatus can help maintain a machine.
Savage’s personalistic probability

In a “small world,” \((S, C)\),
- \(s \in S\) is a way my situation might turn out
- \(f(s) \in C\) is my personal consequence of my action under \(s\)

Savage assumes . . .

1. The existence of complete ranking
2. The independence postulate
3. Value can be purged of belief
4. Belief can be discovered from preference
5. The nontriviality condition
6. The continuity condition
7. The dominance condition

Implication: A person’s preferences among acts can be represented by expected utility relative to a Bayesian prior

Impact: Many subjective Bayesians seek experts to ask about personalistic prior beliefs
The prisoner's dilemma

Two prisoners, P1 and P2, were once colleagues in crime.

Each is offered a reduced sentence for “ratting out” the other.

Payoffs:

<table>
<thead>
<tr>
<th></th>
<th>P1: Remain silent</th>
<th>P1: Rat</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2: Remain silent</td>
<td>(1, 1)</td>
<td>(-1, 2)</td>
</tr>
<tr>
<td>P2: Rat</td>
<td>(2, -1)</td>
<td>(0, 0)</td>
</tr>
</tbody>
</table>

If P1 and P2 act **personally**, each should “rat”.

If P1 and P2 **plan cooperatively**, each should remain silent.
Metrics for fingerprints

Neumann et al.'s metric:
Bayesian hypothesis testing

Hypotheses: $H_0$: same vs $H_1$: different

Bayes factor: Weight of evidence for $H_0$,

$$BF_{01}(Y) = \frac{\pi(Y|H_0)}{\pi(Y|H_1)} = \frac{P[H_0|Y]}{P[H_1|Y]} / \frac{P[H_0]}{P[H_1]}$$

Scales of evidence:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>log_{10} BF</td>
<td>BF</td>
<td>2 log_{e} BF</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>1/2</td>
<td>3.2</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

evidence

- very strong
- strong
- positive
- bare mention
Forensic matching of pattern data

Data: \( X_1(t) \): mark from crime scene
\( Y_1(t) \): mark from person of interest

Are the two marks made by the same person?

Challenges: Sample sizes are very small, \( n_X = n_Y = 1 \)
\( \Rightarrow \) Prior information carries a lot of weight

Preprocessing: Use your favorite metric to translate

\[
\begin{align*}
\text{digitized pattern images} & \quad \text{numerical summary vectors} \\
\quad X_1(t), Y_1(t) & \quad \Rightarrow \quad X_1, Y_1
\end{align*}
\]

Approaches: (1) Develop a model for images
(2) Understand variability in summary vectors
Forensic matching of pattern data

Likelihood: Work with relevant parametric families

\[ X_1 = \theta_X + \epsilon_X \quad \text{with} \quad \epsilon_X | \Sigma_X \sim N(0, \Sigma_X) \]
\[ Y_1 = \theta_Y + \epsilon_Y \quad \text{with} \quad \epsilon_Y | \Sigma_Y \sim N(0, \Sigma_Y) \]

Priors:

\[ H_0 : \pi(\theta_X = \theta_Y, \Sigma_X, \Sigma_Y) \quad \text{vs} \quad H_0 : \pi(\theta_X, \theta_Y, \Sigma_X, \Sigma_Y) \]

Some of what we already know is found in databases

\[ D_S = \text{database for measurement instrument } S \in S \]

Requirements:

- Multiple measurement instruments
- Multiple measurements per individual
Demography

Daponte, Kadane, and Wolfson forecast what the Iraqi Kurdish population from 1977-1990 would have looked like had the repression of the Kurds since 1977 not occurred.

- Knowledge is collected, compiled, and expressed using probability distributions.
- It is rigorously assembled from fertility, mortality, and migration rates, specific to time, age, rural/urban.
- It is based on data from various surveys, censuses, reports, and established model life-tables.
- This process enhances communication among demographers by making beliefs explicit.

Why analyze data?: Data analysis guides a trained community in cooperating to specify prior knowledge.
Forensic matching of pattern data

**Database**

\((D_S : S \in S_X), \text{ like "scene"}\)

\((D_S : S \in S_Y), \text{ like "lab"}\)

\[ \downarrow \]

\(D_X, \text{ at "scene"}\)

\(D_Y, \text{ at "lab"}\)

\[ \downarrow \]

\(X_1, Y_1\)

**Updating**

Prior info for

\(\theta_X, \theta_Y, \Sigma_X, \Sigma_Y\)

\[ \downarrow \]

Update

\(\Sigma_X, \Sigma_Y\)

\[ \downarrow \]

Posterior and Bayes factor

**Goal:** A capability to . . .

- rigorously assemble prior knowledge
- using publicly available databases
- that are discussed and maintained to reflect community knowledge
Putting it together

**Vision:** A (large) database, managed and updated by a trained community, representing prior knowledge about a particular type of pattern data

**Use:** Apply machine learning techniques to create the best metrics for compressing information
Use statistical concepts to design a database for all the information that is needed for inference

**Other conclusions:**
- Not all data analyses are meaningful in the same way, even if they incorporate similar devices and terminology
- Important philosophical shift in the nature of prior knowledge from personalistic to community
Painting by numbers

Dissident artists Vitaly Komar and Alexander Melamid used opinion polls to estimate of the wishes of the *vox populi*

- Beginning late in 1993, they polled 1,001 Americans, regarding preferences as to color, dimensions, settings
- Based on the responses they created the “most wanted” and “least wanted” paintings
- In 1996 they extended into music, creating the “most wanted” and “least wanted” songs

**Why analyze data?:** To put “into question not only the relation between art and ordinary people, and the meaning of ‘the market,’ but also the ambiguity of opinion polls and, by extension, the discordance between the individual and the mass.”
Concrete poetry

In 1965, the “concrete” poet Aram Saroyan wrote a now-famous short poem, which appears as a single misspelled word positioned at the center of the page:

light

Not much more than a point, this poem resembles a single number not unlike one produced in a quantitative inquiry.

What gives this poem meaning? What is the intellectual machinery at play here? Why do people still talk about this poem?

Could the answer be the same as why we analyze data?
Thank You!!
Many short application examples are taken from *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, by Viktor Mayer-Schönberger and Kenneth Cukier.

For more on curve analysis, see


The bullet lands metric is developed in

The fingerprint metric is developed in
References

For good books on the foundations of statistics, especially Bayesian statistics, see

For more on DeFinetti’s representation theorems, especially its history, see

For more on Savage’s personalistic probability, and an important criticism of it, see