Multidimensional Visualization

Maneesh Agrawala

CS 294-10: Visualization
Spring 2011

Last Time:
Exploratory Data Analysis

Topics

Exploratory Data Analysis
Data Diagnostics
Graphical Methods
Data Transformation
Confirmatory Data Analysis
Statistical Hypothesis Testing

Exploratory Analysis:
Effectiveness of Antibiotics

What questions might we ask?

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Streptomycin</th>
<th>Streptomycin</th>
<th>Gentamicin</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Antibiotic megarum</em></td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Bordetella pertussis</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Diphtheria pseudotuberculosis</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Mycobacterium bovis</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Pseudomonas cepacia</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Salmonella enteritidis</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Staphylococcus albus</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Streptococcus faecalis</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td><em>Streptococcus faecalis</em></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Will Burtin, 1951

How do the drugs compare?
How do the bacteria group with respect to antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)
Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009

Common Data Transformations

<table>
<thead>
<tr>
<th>Transform</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalize</td>
<td>$y_i / \sum y_i$ (among others)</td>
</tr>
<tr>
<td>Log</td>
<td>$\log y$</td>
</tr>
<tr>
<td>Power</td>
<td>$y^{1/k}$</td>
</tr>
<tr>
<td>Box-Cox Transform</td>
<td>$(y^\lambda - 1) / \lambda$ if $\lambda \neq 0$</td>
</tr>
<tr>
<td></td>
<td>$\log y$</td>
</tr>
<tr>
<td></td>
<td>if $\lambda = 0$</td>
</tr>
<tr>
<td>Binning</td>
<td>e.g., histograms</td>
</tr>
<tr>
<td>Grouping</td>
<td>e.g., merge categories</td>
</tr>
</tbody>
</table>

Often performed to aid comparison (% or scale difference) or better approx. normal distribution

Lessons

Exploratory Process
1. Construct graphics to address questions
2. Inspect “answer” and assess new questions
3. Repeat!
Transform the data appropriately (e.g., invert, log)
“Show data variation, not design variation”

-Tufte

The Data Set (~200 rows)

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turker ID</td>
<td>String</td>
</tr>
<tr>
<td>Avg. Completion Rate</td>
<td>Number [0,1]</td>
</tr>
</tbody>
</table>

Collected in 2009 by Heer & Bostock.
What questions might we ask of the data?
What charts might provide insight?
Box (and Whiskers) Plot

- Min
- Lower Quartile
- Median
- Upper Quartile
- Max

Turker Completion Percentage

Dot Plot (with transparency to indicate overlap)

Turker Completion Percentage

Dot Plot w/ Reference Lines

Turker Completion Percentage

Histogram (binned counts)

Stem-and-Leaf Plot

Used to compare two distributions; in this case, one actual and one theoretical.

Plots the quantiles (here, the percentile values) against each other.

Similar distributions lie along the diagonal. If linearly related, values will lie along a line, but with potentially varying slope and intercept.

Turker Task Group Completion %
Lessons

Even for “simple” data, a variety of graphics might provide insight. Again, tailor the choice of graphic to the questions being asked, but be open to surprises.

Graphics can be used to understand and help assess the quality of statistical models.

Premature commitment to a model and lack of verification can lead an analysis astray.

Confirmatory Data Analysis

Some Uses of Formal Statistics

What is the probability that the pattern I’m seeing might have arisen by chance?

With what parameters does the data best fit a given function? What is the goodness of fit?

How well do one (or more) data variables predict another?

…and many others.

Example: Heights by Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male / Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (in)</td>
<td>Number</td>
</tr>
</tbody>
</table>
| $\mu_m = 69.4$ | $\sigma_m = 4.69$ | $N_m = 1000$
| $\mu_f = 63.8$ | $\sigma_f = 4.18$ | $N_f = 1000$

Is this difference in heights significant?

In other words: assuming no true difference, what is the prob. that our data is due to chance?
Formulating a Hypothesis

Null Hypothesis ($H_0$): $\mu_m = \mu_f$ (population)
Alternate Hypothesis ($H_a$): $\mu_m \neq \mu_f$ (population)

A statistical hypothesis test assesses the likelihood of the null hypothesis.
What is the probability of sampling the observed data assuming population means are equal? This is called the $p$ value.

Testing Procedure

Compute a test statistic. This is a number that in essence summarizes the difference.
Compute test statistic

\[ Z = \frac{\mu_m - \mu_f}{\sqrt{\sigma^2_m / N_m + \sigma^2_f / N_f}} \]

\[ \mu_m - \mu_f = 5.6 \]

Testing Procedure

Compute a test statistic. This is a number that in essence summarizes the difference.

The possible values of this statistic come from a known probability distribution.

According to this distribution, look up the probability of seeing a value meeting or exceeding the test statistic. This is the \( p \) value.

Lookup probability of test statistic

Normal Distribution
\( \mu = 0, \sigma = 1 \)
\( Z ~ N(0, 1) \)

95% of Probability Mass

\( p < 0.05 \)

\( Z = 0.2 \)
\( Z > 1.96 \)

Statistical Significance

The threshold at which we consider it safe (or reasonable?) to reject the null hypothesis.

If \( p < 0.05 \), we typically say that the observed effect or difference is statistically significant.

This means that there is a less than 5% chance that the observed data is due to chance.

Note that the choice of 0.05 is a somewhat arbitrary threshold (chosen by R. A. Fisher)

Common Statistical Methods

<table>
<thead>
<tr>
<th>Question</th>
<th>Data Type</th>
<th>Parametric</th>
<th>Non-Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do data distributions have different &quot;centers&quot;? (aka &quot;location&quot; tests)</td>
<td>&gt; 2 uni. data</td>
<td>t-Test</td>
<td>Mann-Whitney U</td>
</tr>
<tr>
<td>Are observed counts significantly different?</td>
<td>&gt; 2 multi. data</td>
<td>ANOVA</td>
<td>Kruskal-Wallis</td>
</tr>
<tr>
<td>Are two vars related?</td>
<td>Counts in categories</td>
<td>MANOVA</td>
<td>Median Test</td>
</tr>
<tr>
<td>Do 1 (or more) variables predict another?</td>
<td>2 variables</td>
<td>chi-squared</td>
<td>( z^2 ) (chi-squared)</td>
</tr>
<tr>
<td>Continuous ( X )</td>
<td>Binary</td>
<td>Pearson coeff</td>
<td>Rank correl</td>
</tr>
<tr>
<td>Logistic regression</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

Exploratory analysis may combine graphical methods, data transformations, and statistics. Use questions to uncover more questions. Formal methods may be used to confirm, sometimes on held-out or new data.

Assignment 2: Exploratory Data Analysis

Use existing software to formulate & answer questions

First steps
- Step 1: Pick a domain
- Step 2: Pose questions
- Step 3: Profile data
- Iterate

Create visualizations
- Interact with data
- Refine your questions
- Tableau

Make wiki notebook
- Keep record of all steps you took to answer the questions

Due before class on Feb 14, 2011

Multidimensional Visualization

Visual Encoding Variables

Position  Length  Area  Volume  Value  Texture  Color  Orientation  Shape

~8 dimensions?

Small Multiples [from Wills 95]

how long in majors
avg assists vs avg putouts (fielding ability)
distribution of positions played

select high salaries
avg career HRs vs avg career hits (batting ability)
Scatterplot Matrix (SPLOM)

Scatter plots enabling pair-wise comparison of each data dimension.

Dimensional Projection

http://www.ggobi.org/

Chernoff Faces (1973)

Insight: We have evolved a sophisticated ability to interpret facial expression

Idea: Map data variables to facial features

Question: Do we process facial features in an uncorrelated way? (i.e., are they separable?)

This is just one example of nD "glyphs"

Visualizing Multiple Dimensions

Strategies

- Avoid “over-encoding”
- Use space and small multiples intelligently
- Reduce the problem space
- Use interaction to generate relevant views

There is rarely a single visualization that answers all questions. Instead, the ability to generate appropriate visualizations quickly is key

Tableau / Polaris
Tableau Research at Stanford: “Polaris” by Stolte and Hanrahan.

Tableau

Data Display

Data Model

Tableau demo

The dataset:
- Federal Elections Commission Receipts
- Every Congressional Candidate from 1996 to 2002
- 4 Election Cycles
- 9216 Candidacies

Data Set Schema
- Year (Qi)
- Candidate Code (N)
- Candidate Name (N)
- Incumbent / Challenger / Open-Seat (N)
- Party Code (N) [1=Dem,2=Rep,3=Other]
- Party Name (N)
- Total Receipts (Qr)
- State (N)
- District (N)

This is a subset of the larger data set available from the FEC, but should be sufficient for the demo

Hypotheses?

What might we learn from this data?
- Correlation between receipts and whether elected?
- Do receipts increase over time?
- Which states spend the most?
- Which party spends the most?
- Margin of victory vs. amount spent?
- Amount spent between competitors?

Hypotheses?

What might we learn from this data?
- Has spending increased over time?
- Do democrats or republicans spend more money?
- Candidates from which state spend the most money?
**Polaris/Tableau Approach**

Insight: simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Recently: automate visualization design (ShowMe – Like APT)

---

**Specifying Table Configurations**

Operands are names of database fields

Each operand interpreted as a set (…)

Quantitative and Ordinal fields treated differently

Three operators:
- concatenation (+)
- cross product (x)
- nest (/)

---

**Table Algebra: Operands**

**Ordinal fields**: interpret domain as a set that partitions table into rows and columns

Quarter = {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} →

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Coffee</th>
<th>Espresso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qtr1</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>Qtr2</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>Qtr3</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>Qtr4</td>
<td>70</td>
<td>80</td>
</tr>
</tbody>
</table>

**Quantitative fields**: treat domain as single element set and encode spatially as axes

Profit = {Profit[-10,650]} →

Profit = Profit[-10,650]

---

**Concatenation (+) Operator**

Ordered union of set interpretations

Quarter + Product Type = {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} + {(Coffee), (Espresso)}

Profit + Sales = {(Profit[-310,620]),(Sales[0,1000])}

---

**Cross (x) Operator**

Cross-product of set interpretations

Quarter x Product Type = {(Qtr1,Coffee), (Qtr1, Tea), (Qtr2, Coffee), (Qtr2, Tea), (Qtr3, Coffee), (Qtr3, Tea), (Qtr4, Coffee), (Qtr4, Tea)}

Product Type x Profit =

---

**Nest (/) Operator**

Cross-product filtered by existing records

Quarter x Month creates twelve entries for each quarter. i.e., (Qtr1, December)

Quarter / Month creates three entries per quarter based on tuples in database (not semantics)
Ordinal - Ordinal

Quantitative - Quantitative

Ordinal - Quantitative

Querying the Database