Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

<table>
<thead>
<tr>
<th>Set A</th>
<th></th>
<th>Set B</th>
<th></th>
<th>Set C</th>
<th></th>
<th>Set D</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
<td>13</td>
<td>12.74</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.11</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
</tr>
</tbody>
</table>

Summary Statistics
- $u_x = 9.0$, $s_x = 3.317$
- $u_y = 7.5$, $s_y = 2.03$

Linear Regression
- $y = 3 + 0.5x$
- $R^2 = 0.67$

Anscombe 1973
Antibiotic Effectiveness

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Streptomycin</th>
<th>Neomycin</th>
<th>Gram Staining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerobacter aerogenes</td>
<td>870</td>
<td>1</td>
<td>1.6</td>
<td>negative</td>
</tr>
<tr>
<td>Brevibacter subtilis</td>
<td>8</td>
<td>2</td>
<td>0.02</td>
<td>negative</td>
</tr>
<tr>
<td>Brevibacter anthracis</td>
<td>0.605</td>
<td>0.01</td>
<td>0.007</td>
<td>positive</td>
</tr>
<tr>
<td>Diplococcus pneumoniae</td>
<td>0.605</td>
<td>11</td>
<td>10</td>
<td>positive</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>100</td>
<td>0.4</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td>Eubacterium aerogenes</td>
<td>820</td>
<td>1.2</td>
<td>1</td>
<td>negative</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>830</td>
<td>5</td>
<td>3</td>
<td>negative</td>
</tr>
<tr>
<td>Pasteurella multocida</td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>850</td>
<td>2</td>
<td>0.4</td>
<td>negative</td>
</tr>
<tr>
<td>Salmonella (Enteritidis)</td>
<td>0.4</td>
<td>0.4</td>
<td>0.008</td>
<td>negative</td>
</tr>
<tr>
<td>Salmonella (cholerae)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.09</td>
<td>negative</td>
</tr>
<tr>
<td>Staphylococcus albus</td>
<td>0.007</td>
<td>0.1</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td>Staphylococcus epidermidis</td>
<td>0.001</td>
<td>0.01</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td>Staphylococcus warneri</td>
<td>0.01</td>
<td>0.01</td>
<td>0.001</td>
<td>positive</td>
</tr>
</tbody>
</table>

Will Burtin, 1951

How do the drugs compare?

How do the bacteria group w.r.t. resistance?

Do different drugs correlate?

Mike Bostock, CS448B Winter 2009

American Scientist, 2009
Transforming data
How well does the curve fit the data?

Plot the Residuals
Plot vertical distance from best fit curve
Residual graph shows accuracy of fit

Multiple Plotting Options
- Plot model in data space
- Plot data in model space

Choropleth maps of cancer deaths in Texas.
One plot shows a real data set. The others are simulated under the null hypothesis of spatial independence.
Can you spot the real data? If so, you have some evidence of spatial dependence in the data.
Multidimensional Visualization

Visual Encoding Variables
- Position
- Length
- Area
- Volume
- Value
- Texture
- Color
- Orientation
- Shape

~8 dimensions?

Example: Coffee Sales
Sales figures for a fictional coffee chain:
- Sales: Q-Ratio
- Profit: Q-Ratio
- Marketing: Q-Ratio
- Product Type: N (Coffee, Espresso, Herbal Tea, Tea)
- Market: N (Central, East, South, West)
Encode “Sales” (Q) and “Profit” (Q) using Position

Encode “Product Type” (N) using Hue

Encode “Market” (N) using Shape

Encode “Marketing” (Q) using Size
Trellis Plots

A trellis plot subdivides space to enable comparison across multiple plots. Typically nominal or ordinal variables are used as dimensions for subdivision.

Scatterplot Matrix (SPLOM)

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

- 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)

Dimensionality Reduction

Linking Assists to Positions

Principal Component Analysis

1. Mean-center the data.
2. Find basis vectors that maximize the data variance.
3. Plot the data using the top vectors.

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.

Parallel Coordinates
**The Multidimensional Detective**

The Dataset:
- Production data for 473 batches of a VLSI chip
- 16 process parameters:
  - X1: The yield, % of produced chips that are useful
  - X2: The quality of the produced chips (speed)
  - X3 … X12: 10 types of defects (zero defects shown at top)
  - X13 … X16: 4 physical parameters

The Objective:
Raise the yield (X1) and maintain high quality (X2)

**Inselberg’s Principles**

1. Do not let the picture scare you
2. Understand your objectives
   - Use them to obtain visual cues
3. Carefully scrutinize the picture
4. Test your assumptions, especially the “I am really sure of’s”
5. You can’t be unlucky all the time!

Look for batches with *nearly* zero defects (9/10)
Most of these have low yields ➔ defects OK.
Notice that X6 behaves differently.
Allow 2 defects, including X6 → best batches.

![Radar Plot / Star Graph](image)

“Parallel” dimensions in polar coordinate space
Best if same units apply to each axis

**Tableau / Polaris**

**Polaris**
Research at Stanford by Stolte, Tang, and Hanrahan.
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

Hypotheses?

What might we learn from this data?
• ??
Hypotheses?
What might we learn from this data?
Correlation between receipts and winners?
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?

Assignment 2: Exploratory Data Analysis
Use visualization software (Tableau) to form & answer questions
First steps:
- Step 1: Pick domain & data
- Step 2: Pose questions
- Step 3: Profile the data
- Iterate as needed
Create visualizations
- Interact with data
- Refine your questions
Make wiki notebook
- Keep record of your analysis
- Prepare a final graphic and caption

Due by end-of-day
Tuesday, October 18

Tableau Demo

Polaris/Tableau Approach
Insight: can simultaneously specify both database queries and visualization
Choose data, then visualization, not vice versa
Use smart defaults for visual encodings
More recently: automate visualization design
Specifying Table Configurations

Operands are the 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>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>9032</td>
<td>10176</td>
<td>11320</td>
<td>90225</td>
</tr>
</tbody>
</table>

Quantitative fields: treat domain as single element set and encode spatially as axes:
Profit = \{(Profit[-410,650])\} →

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)\}

Business Insights

- Analyze sales and profit across different quarters and product types.
- Identify trends and make informed decisions based on data analysis.
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)
Visualizing Multiple Dimensions

Strategies

- Start by visualizing individual dimensions
- Avoid “over-encoding”
- Use space and small multiples intelligently
- 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.