Correlation and Causation

Privacy
Correlation and Causation

Analyzing data with two (or more) measures $X$ and $Y$

- Height and shoe size
- Grades and SAT scores
- Education level and starting salary
- Temperature and ice cream sales

- **Correlation (informal):** The values of $X$ and $Y$ tend to be interdependent

- **Causation (informal):** $X$’s value tends to influence $Y$’s value
Why Do We Care?

- **Discoveries in the medical domain**
  - Patients with elevated X also tend to have elevated Y
  - Taking drug X tends to make symptom Y subside

- **Discoveries in the political domain**
  - Voters who approve of X tend to also approve of Y
  - Voter turnout is weather-dependent

- **Discoveries in the advertising domain**
  - Larger fonts tend to cause more click-throughs
  - More purchases are made in the evening
Categorical versus Numeric Values

- **Categorical values:** unordered categories
  - color, weather, dormitory

- **Numeric values:** ordered values
  - height, price, time, age, SAT score
  - May be discrete or continuous

- **Ordinal values:** categories that can be ordered
  - movie rating, letter grade, education level
  - But differences may not be on a meaningful scale

We’ll stick to ordered for now
Positive Correlation

Two measures $X$ and $Y$

- When $X$ is higher $Y$ tends to be higher
- When $X$ is lower $Y$ tends to be lower
- When $Y$ is higher $X$ tends to be higher
- When $Y$ is lower $X$ tends to be lower

Examples

- $X =$ height, $Y =$ shoe size
- $X =$ grades, $Y =$ SAT scores

Notation (mine): $X \approx Y$
Positive Correlation by Causation

Two measures X and Y

* X being higher causes Y to be higher
* X being lower causes Y to be lower

Examples

* X = education, Y = starting salary
* X = temperature, Y = ice cream sales

Notation (mine): $X \rightarrow Y$
Correlation due to Hidden Causation

Correlation can be the result of causation from a hidden “confounding variable”

\[ \text{X } \approx \text{ Y because there’s a hidden Z such that } \]
\[ Z \rightarrow X \text{ and } Z \rightarrow Y \]

Homeless population \( \approx \) crime rate
Confounding variable: unemployment

Forgetfulness \( \approx \) near-sightedness
Confounding variable: age
Negative Correlation by Causation

Two measures $X$ and $Y$

- $X$ being higher causes $Y$ to be lower
- $X$ being lower causes $Y$ to be higher

Examples

- $X = \text{car weight}, \ Y = \text{gas mileage}$
- $X = \text{class absences}, \ Y = \text{final grade}$
Negative Correlation without Causation

Two measures X and Y

- When X is higher Y tends to be lower
- When X is lower Y tends to be higher
- When Y is higher X tends to be lower
- When Y is lower X tends to be higher

Examples

X = ice cream sales, Y = hot chocolate sales
X = years of schooling, Y = years in jail

Confounding variables?
Is There Such a Thing as Pure Correlation?

Correlation without causation: usually a confounding variable lurking somewhere

\[ X = \text{height}, \ Y = \text{shoe size} \]
\[ X = \text{grades}, \ Y = \text{SAT scores} \]

What about the “spurious correlations”?
1) Want to know when things are correlated
2) But should not assume one causes the other
   “Correlation does not imply causation”

Next:
• Determining correlation
• Determining if there’s causation
Determining Correlation

X and Y both ordered: scatterplot
Determining Correlation

X and Y both ordered: scatterplot
Determining Correlation

X categorical, Y ordered: bar graph

Average temperature per region

Must use average or median, not sum
Determining Correlation

X categorical, Y ordered: bar graph
Determining Correlation

X and Y both categorical: table

<table>
<thead>
<tr>
<th></th>
<th>Declared</th>
<th>Undeclared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Sophomore</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Junior</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Senior</td>
<td>24</td>
<td>1</td>
</tr>
</tbody>
</table>

(Ignore that year is ordinal)

Without correlation expect “evenly divided values”
Determining Correlation

**X and Y both categorical: table**

<table>
<thead>
<tr>
<th></th>
<th>Declared</th>
<th>% of row</th>
<th>Undeclared</th>
<th>% of row</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>0</td>
<td>0%</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>Sophomore</td>
<td>2</td>
<td>18%</td>
<td>9</td>
<td>82%</td>
</tr>
<tr>
<td>Junior</td>
<td>3</td>
<td>37%</td>
<td>5</td>
<td>63%</td>
</tr>
<tr>
<td>Senior</td>
<td>24</td>
<td>96%</td>
<td>1</td>
<td>4%</td>
</tr>
</tbody>
</table>

If different rows have different relative percentages, values are correlated
## Determining Correlation

### X and Y both categorical: table

<table>
<thead>
<tr>
<th></th>
<th>Declared</th>
<th>% of column</th>
<th>Undeclared</th>
<th>% of column</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>0</td>
<td>0%</td>
<td>2</td>
<td>12%</td>
</tr>
<tr>
<td>Sophomore</td>
<td>2</td>
<td>7%</td>
<td>9</td>
<td>53%</td>
</tr>
<tr>
<td>Junior</td>
<td>3</td>
<td>10%</td>
<td>5</td>
<td>29%</td>
</tr>
<tr>
<td>Senior</td>
<td>24</td>
<td>83%</td>
<td>1</td>
<td>6%</td>
</tr>
</tbody>
</table>

If different columns have different relative percentages, values are correlated.
Determining Causation

No simple data analysis technique

1) “Hill’s Criteria”
2) Running experiments
3) One new statistical approach
Hill’s Criteria (slightly adapted)

Strength - of correlation between X and Y
Consistency - of correlation across different datasets
Specificity - no other likely explanation for correlation
Temporality - Y occurs after X
Plausibility - there’s a reason for causation
Coherence - consistent with related theories
Experimental Validation

To determine if $X \rightarrow Y$

- Create “experimental group” with $X=value$
- Create “control group” with different $X$ values
- Ensure no other distinctions between two groups
- See if experimental vs. control $\approx Y$

Works for things like drug therapy, advertising font size

Less useful for things like weather, crime rate, forgetfulness
New Statistical Approach

“Additional noise model”

For cases when it’s known either $X \rightarrow Y$ or $Y \rightarrow X$, i.e., no confounding variable

- When there are variations (“noise”) in $X$ then there are also variations in $Y$
- But not necessarily vice-versa

- Conclude $X \rightarrow Y$
Historical Example

Smoking and Lung Cancer, 1950’s

- Strong correlation observed between smoking and lung cancer
- Tobacco proponents implicated confounding variables (e.g., pollution, occupation, genetic predisposition)
- Experimental validation not feasible
- Large-scale data analysis (Big Data!) concluded causation, both using Hill’s criteria and eliminating proposed confounding variables
Privacy

1) Individual data collected covertly

2) Individual data collected legally, but perhaps used unethically

3) Individual data deduced from “anonymous” public data

4) Invisible opt-in
Individual Data Collected Covertly

- National Security Agency collecting phone records and internet traffic of most Americans
- Revealed by whistleblower Edward Snowden
- NSA argued “metadata” does not invade privacy
  But easy to detect medical conditions, psychological conditions, criminal activity
- Since ruled illegal
Data Collected Legally, Used Unethically?

Individual “digital footprints” are large and revealing, often not private
- web browsing, online and in-store purchases,
- social media posts, online ratings & reviews,
- search queries, photos, check-ins, mobile phone locations, text messages, ...

Target pregnancy predictor
- Machine learning from customer purchases prior to joining baby gift registry
  (Solution: hide promotions inside booklet)

Facebook “Beacon”
- Diamond ring purchase broadcast, lawsuit followed
Deducing Individual Data

Even without personal identifiers, can be easy to identify a person

5-digit zip + gender + date of birth uniquely identifies 87% of U.S. population

Ex: “Anonymous” patient data distributed by insurance commission

Data analysts identified Massachusetts governor
Correlation and Causation

Privacy