Uncovering Hidden Insights with Data Driven Hypothesis Testing
Key Features

• Represent knowledge in the form of comparison, i.e., hypotheses… a more nature way
• Formulation of hypotheses are automatic and data-driven, instead of expert-driven
• Can detect hidden phenomena, e.g., Simpson’s paradox
• Multiple test corrections and robust statistics are employed to ensure the validness of findings
• Useful for a wide range of applications, e.g., business intelligence, census analysis, …
Knowledge Representation & Discovery via Comparison
Comparison is a natural way to represent knowledge

How about NOW?
“Comparison”

- Statistically defined as Hypothesis

- Hypothesis vs Rule

**Rule-based Knowledge**

For *female, <30-year old* customers, 38% of them respond positively to our promotion.

**BUT** is 38% a good or bad response rate?
“Comparison”: Hypothesis vs Rule

**Hypothesis-based Knowledge**

The positive response rate of *female, <30-year old* customers (38%) is *significantly* \( p<0.05 \) higher than the average rate (16%).

**NOW** we can fully appreciate the meaning of 38%!
Advantages

• Consolidates info from multiple rules
• Comparing groups that share similarities but have different behaviors due to one differentiating factor

<table>
<thead>
<tr>
<th>Rule-based Knowledge</th>
<th>Hypothesis-based Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, &lt;30-year-old, gold-card customers spend $500 per month on our products.</td>
<td>Among female, &lt;30-year-old customers, gold-card holders’ monthly spending ($500) is significantly (p &lt; 0.05) higher than silver-card holders ($300).</td>
</tr>
<tr>
<td>Female, &lt;30-year-old, silver-card customers spend $300 per month on our products.</td>
<td></td>
</tr>
</tbody>
</table>
Advantages

- More reliable findings with statistical reasoning

Observations
Sales on November is $300K; sales on December is $400K

Question
Is this sales increase statistically significant?

Scenario 1: Non-significant
Scenario 2: Significant
Hypothesis Testing

• Null Hypothesis
  Monthly spending of gold-card holders, $\mu_g$, are the same as silver-card holders, $\mu_s$, i.e. $\mu_g = \mu_s$

• Alternative Hypothesis: $\mu_g \neq \mu_s$

• Reject the null hypothesis, if and only if
  – $P < \alpha$
  – Significance level, $\alpha$, --- the false positive rate

• $P$-value
  – The probability, assuming the null hypothesis is true, of observing a result at least as extreme as the observed statistics
Expert-Driven Hypothesis Analysis

- Conventional hypothesis analysis is expert-driven
  - Expert dependent
  - Rich prior knowledge is required
  - Already have a clear question in mind
  - Often not applicable in real applications
Data-Driven Hypothesis Analysis

• iDIG is Data-driven
  – Minimum expert input is required
  – Suitable for analysis scenarios, where
    • Little prior knowledge is available
    • Investigating question is not clear yet
    • Exploratory studies
Formulation of Hypothesis

• 3 major components
  – Target Attribute:
    • attribute of interest
  – Context Attributes:
    • similarities between comparing groups
  – Comparing Attribute:
    • attribute that differentiate comparing groups
Formulation of Hypothesis: Target Attribute

• **Attribute of interest/attribute to be investigated**
• **E.g**
  – Monthly spending
  – Response rate to product/promotion
  – Performance of drugs
  – Grades of students, etc
• **Can be categorical or continuous**
Formulation of Hypothesis: Context Attributes

- Attributes that defines similarities among comparing groups
- Defines a sub-population of interest
- Must be categorical
- Example:
  - Female, <30-year-old customers
Formulation of Hypothesis: Context Attributes

- Attribute that differentiate the comparing groups
- Single attribute
  - To avoid confusion due to confounding factors
- Must be categorical
- Example
  - Card type: gold-card vs silver-card
Formulation of Hypothesis

<table>
<thead>
<tr>
<th>Context Attributes</th>
<th>Comparing Attribute: Card-type</th>
<th>Target Attribute: Monthly Spending</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, &lt;30-year-old</td>
<td>Gold-card</td>
<td>$500</td>
</tr>
<tr>
<td></td>
<td>Silver-card</td>
<td>$300</td>
</tr>
</tbody>
</table>

Formulated Hypothesis

Among *female, <30-year-old* customers, *gold-card* holders’ monthly spending (*$500*) is significantly higher than *silver-card* holders’ (*$300*).
Discovery of Hypothesis
Hypothesis Testing

• Categorical target attribute
  – $\chi^2$ test
    • Computational efficient
    • Approximation test --- tend to underestimate p-value
    • Some constrains on the data
  – Fisher’s exact test
    • Exact test --- accurate p-value estimation
    • Computationally more expensive
    • Impractical for more than 2 groups of comparisons
Hypothesis Testing

- **Continuous target attribute**
  - T-test
    - Computational efficient
    - Normality is necessary
  - Wilcoxon rank-sum test
    - Computationally more expensive
    - Normality is not necessary
    - BUT, data’s distribution must be symmetrical
  - Permutation test
    - Most reliable,
    - BUT computationally very expensive
Important Features

• Detection of Simpson’s Paradox

• Correction for Multiple Test Effect

• Employment of Robust Statistics
  – Protection against outliers and artifacts

• Visualization of Knowledge
Simpson’s Paradox

• Charing et al 1986
  – Study of treatments for kidney stones

<table>
<thead>
<tr>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment A</td>
</tr>
<tr>
<td>Treatment B</td>
</tr>
</tbody>
</table>

– Seems Treatment B is more effective
– **BUT**, if we break down the patients in two groups: small stones & big stones

<table>
<thead>
<tr>
<th>Small Stone</th>
<th>Big Stone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment A</td>
<td>81/87 (93%)</td>
</tr>
<tr>
<td>Treatment B</td>
<td>234/270 (87%)</td>
</tr>
</tbody>
</table>

– Treatment A performs better in both subgroup
– A paradox!
Multiple Test Effect

• Occurs when one considers multiple comparison simultaneously
  – i.e., when one tests multiple hypotheses over the same set of data
• Example:

  “Suppose we consider the safety of a new drug in terms of its side effects. As more and more types of side effects are considered, sooner or later one will find at least one side effect, where the new drug is “significantly” higher than the existing one.”
Multiple Test Effect

- **Give single hypothesis significance level, \( \alpha \)**
  - The chance of rejecting a null hypothesis when it is true is controlled to be at most \( \alpha \) --- Type I error or false positive
  - If \( m \) independent hypotheses were tested, the experiment-wide significance level, \( \alpha_w \), is

\[
\alpha_w = 1 - (1 - \alpha_s)^m
\]

where \( \alpha_s \) is the significance level for single hypothesis

- i.e, if 100 hypotheses are tested on the same data, where \( \alpha_s = 5\% \), then \( \alpha_w = 99.4\% \)
Multiple Test Effect

- **Bonferroni correction**
  - Simple but conservative
  
  \[ \alpha_s = \alpha_w / m \]

  where \( \alpha_s \) and \( \alpha_w \) are the corrected and targeted significance level, and \( m \) is the number of tested hypotheses.

- **Sidak correction**
  - Better correction

  \[ \alpha_w = 1 - (1 - \alpha_s)^m \]

  \[ \alpha_s = 1 - (1 - \alpha_w)^{1/m} \]
Robust Statistics
Protection from Outliers & Artifacts

- **Effect of outliers and artifacts**

<table>
<thead>
<tr>
<th>Customer</th>
<th>Rating for Pizza A</th>
<th>Rating for Pizza B</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.4</td>
<td>20.2</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>24.2</td>
<td>16.9</td>
<td>7.3</td>
</tr>
<tr>
<td>3</td>
<td>25.4</td>
<td>18.5</td>
<td>6.9</td>
</tr>
<tr>
<td>4</td>
<td>21.4</td>
<td>17.3</td>
<td>4.1</td>
</tr>
<tr>
<td>5</td>
<td>20.2</td>
<td>15.5</td>
<td>4.7</td>
</tr>
<tr>
<td>6</td>
<td>21.5</td>
<td>18.5</td>
<td>3.0</td>
</tr>
</tbody>
</table>

**Sample Mean**: 4.36
**Sample Variance**: 2.62
**Estimated test statistics**: 4.08

**p-value (2 tail, paired t-test)**: 0.01

<table>
<thead>
<tr>
<th>Customer</th>
<th>Rating for Pizza A</th>
<th>Rating for Pizza B</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.4</td>
<td>20.2</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>24.2</td>
<td>16.9</td>
<td>7.3</td>
</tr>
<tr>
<td>3</td>
<td>25.4</td>
<td>18.5</td>
<td>6.9</td>
</tr>
<tr>
<td>4</td>
<td>21.4</td>
<td>17.3</td>
<td>4.1</td>
</tr>
<tr>
<td>5</td>
<td>20.2</td>
<td>15.5</td>
<td>4.7</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>18.5</td>
<td>196.5</td>
</tr>
</tbody>
</table>

**Sample Mean**: 36.62
**Sample Variance**: 78.4
**Estimated test statistics**: 1.14
**p-value (2 tail, paired t-test)**: 0.3
Robust Statistics
Protection from Outliers & Artifacts

- Robust statistics
  - Mute or down-weigh potential outliers and artifacts

<table>
<thead>
<tr>
<th>Customer</th>
<th>Rating for Pizza A</th>
<th>Rating for Pizza B</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.4</td>
<td>20.2</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>24.2</td>
<td>16.9</td>
<td>7.3</td>
</tr>
<tr>
<td>3</td>
<td>25.4</td>
<td>18.5</td>
<td>6.9</td>
</tr>
<tr>
<td>4</td>
<td>21.4</td>
<td>17.3</td>
<td>4.1</td>
</tr>
<tr>
<td>5</td>
<td>20.2</td>
<td>15.5</td>
<td>4.7</td>
</tr>
<tr>
<td>6</td>
<td>21.5</td>
<td>18.5</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Sample Mean: 4.36
Sample Variance: 2.62
Estimated test statistics: 4.08
p-value (2 tail, paired t-test): 0.01

<table>
<thead>
<tr>
<th>Customer</th>
<th>Rating for Pizza A</th>
<th>Rating for Pizza B</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.4</td>
<td>20.2</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>24.2</td>
<td>16.9</td>
<td>7.3</td>
</tr>
<tr>
<td>3</td>
<td>25.4</td>
<td>18.5</td>
<td>6.9</td>
</tr>
<tr>
<td>4</td>
<td>21.4</td>
<td>17.3</td>
<td>4.1</td>
</tr>
<tr>
<td>5</td>
<td>20.2</td>
<td>15.5</td>
<td>4.7</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>18.5</td>
<td>196.5</td>
</tr>
</tbody>
</table>

Sample Mean: 36.62
Sample Variance: 78.4
Estimated test statistics: 1.14
p-value (2 tail, paired t-test): 0.3
Visualization
Capabilities

- Uncover groups that behave differently from the global norm
- Discover groups that share considerable similarities but have significantly different behaviors due to some differentiating factor
- Analyze change of behavior before and after certain events, such as promotion campaigns
- Detect and trace change of behavior over time
  - Dynamic version of iDIG (under-development)
Potential Applications

- **Business Intelligence**
  - Customer segmentation/profiling
  - Customer loyalty study
  - Campaign analysis

- **Medical Studies**
  - Drug effectiveness evaluation
  - Epistasis study
  - Clinical data analysis

- **Expert Knowledge Extraction**
Features of iDIG

• **Data-driven** knowledge discovery tool

• **Represent and discovery knowledge via comparison** (hypothesis analysis)

• **Reliability** of findings are ensured with statistical testing and reasoning

• **Can be customized for a wide range of applications**
Contact

Dr Feng Mengling
mfeng@i2r.a-star.edu.sg

Dr Liu Guimei
liugm@comp.nus.edu.sg

Prof Wong Limsoon
wongls@comp.nus.edu.sg

• Supported in part by A*STAR grants SERC 072 101 0016 and SERC 102 101 0030