Sampling for Discovering the Outstanding #1 Data Insights

Brandon Lockhart

November 29, 2019
In 2018, Google conducted a pay equity analysis and found that in a large job code (Level 4 Software Engineer) men received less discretionary funds than women.¹

<table>
<thead>
<tr>
<th>Job title</th>
<th>Sex</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 4 Software Engineer</td>
<td>Female</td>
<td>$120,000</td>
</tr>
<tr>
<td>Level 4 Software Engineer</td>
<td>Female</td>
<td>$130,000</td>
</tr>
<tr>
<td>Level 4 Software Engineer</td>
<td>Male</td>
<td>$115,000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

This audit resulted in $9.7 million in adjustments to a total of 10,677 employees.

¹https://www.blog.google/inside-google/company-announcements/ensuring-we-pay-fairly-and-equitably/
Consider a drug for lowering blood pressure. For each patient prescribed this drug, attributes are recorded including `weight_class`, `sex`, `smokes_cigarettes (Y/N)`, `age_group`, etc. Also, the patient’s decrease in blood pressure is recorded.

Suppose the average blood pressure decrease when grouped by `sex`, `age_group`, and `smokes_cigarettes` is significantly higher in one group than the rest:

<table>
<thead>
<tr>
<th>sex</th>
<th>age_group (years)</th>
<th>smokes_cigarettes</th>
<th>average decrease in blood pressure (mmHg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>25-44</td>
<td>No</td>
<td>12</td>
</tr>
<tr>
<td>Female</td>
<td>45-64</td>
<td>No</td>
<td>8</td>
</tr>
<tr>
<td>Male</td>
<td>25-44</td>
<td>No</td>
<td>7.7</td>
</tr>
<tr>
<td>Male</td>
<td>25-44</td>
<td>No</td>
<td>7.5</td>
</tr>
<tr>
<td>Female</td>
<td>25-44</td>
<td>Yes</td>
<td>7.4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Consider a multidimensional dataset $\mathcal{R}(\mathcal{D}, M)$ where $\mathcal{D} = \{D_1, \ldots, D_d\}$ is a set of **dimensions** (columns containing values that can be grouped over) and $M$ is a **measure** (column containing numerical values which can be aggregated over).

For each set $S \subseteq \mathcal{D}, S \neq \emptyset$, query

\[
\text{SELECT AVG}(M) \\
\text{FROM } \mathcal{R} \\
\text{GROUP BY } S;
\]

Use sampling to find queries that satisfy the **Outstanding #1 condition**: when one group in a resulting query has a value that is significantly higher than all other groups.
Need to bound the error on the estimate of the mean in each group.

Try using Hoeffding’s bound:

**Confidence interval using Hoeffding’s inequality**

Let $X_1, \ldots, X_n$ be independent random variables bounded by the interval $[a, b]$, and let $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.

When $n \geq \frac{(b-a)^2 \log(2/\alpha)}{2\epsilon^2}$, with probability greater than $(1 - \alpha)$,

$$\bar{X} \in (E[\bar{X}] - \epsilon, E[\bar{X}] + \epsilon).$$

The problem is finding $a$ and $b$ for each group; the computation to find $a$ and $b$ is as expensive as simply calculating the mean in that group.
Second approach: normal confidence intervals

- Assume $X_1, \ldots, X_n$ are an iid sample drawn from a normal distribution, $\bar{X} = \sum_{i=1}^{n} X_i$ is the sample mean and 
  $\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$ is the sample standard deviation.
- $\bar{X}$ is an unbiased estimator of the population mean.
- For each group, calculate the 95% confidence interval for the estimate of the mean as
  \[ \left( \bar{X} - 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right). \]
First approach: SRSWOR

- Take a sample from the dataset and mine for insights on that sample.
- Major problem is that there is no guarantee to see a sample from every group.

![Average housing prices](image)

**Figure:** A dataset recording housing prices may contain very few mansions.

- Confidence intervals could be computed on very small sample sizes, and it is unknown whether a small sample is the entire group.
Second approach: stratified sampling

- Sample from each group
- Too expensive to perform for every set of dimensions
- Instead, group by all dimensions, and sample from each resulting group
- Guarantees a sample in each group

<table>
<thead>
<tr>
<th>Field</th>
<th>Degree</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computing Science</td>
<td>Bachelor</td>
<td>$100,000</td>
</tr>
<tr>
<td>Computing Science</td>
<td>... (2,000)</td>
<td>...</td>
</tr>
<tr>
<td>Computing Science</td>
<td>Bachelor</td>
<td>$100,000</td>
</tr>
<tr>
<td>Computing Science</td>
<td>PhD</td>
<td>$250,000</td>
</tr>
<tr>
<td>Computing Science</td>
<td>... (50)</td>
<td>...</td>
</tr>
<tr>
<td>Computing Science</td>
<td>PhD</td>
<td>$250,000</td>
</tr>
<tr>
<td>Engineering</td>
<td>Bachelor</td>
<td>$40,000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Average salary $100,000
Average salary $250,000

When subsequently grouping by a subset of dimensions, the sample of resulting groups may not be iid.
Combination sampling method used - SRSWOR and stratified sampling

Select a simple random sample of the data. Also, sample from each group after grouping by all dimensions.

- Samples will exist in each group
- Simple random sample could mitigate the iid assumption violation
Outstanding #1 condition: Let $lb_1$ be the lower bound on the confidence interval of the largest estimate of the mean over all groups, and let $ub^*$ be the largest upper bound on the confidence intervals for all other groups. Let $\gamma > 1$ be a threshold. The Outstanding #1 condition is satisfied if

$$lb_1 \geq \gamma ub^*.$$
Experiments - synthetic data

- Parameters: dataset size, number of dimensions, number of categories per dimension, distribution of data values. Also need to set strata sample size and uniform sample size.
- On the basis of my experiments, I found that a strata sample size of 50, and a uniform sample size of 1% often performed well and did not require excessive data when the number of categories per dimension was small.

Figure: Example running times on synthetic data.
Running times from three real datasets:

1. **UK Housing Prices Paid** (≈ 22M rows). 6 dimensions, measure is housing price.
2. **London Crime Data** (≈ 13M rows). 4 dimensions, measure is count of crimes per category per month.
3. **Loans** (≈ 2.3M rows). 6 dimensions, measure is loan request amount.
The table below denotes the accuracy of discovering the Outstanding #1 insights for three sampling approaches. All sampling methods resulted in the same sample size.

<table>
<thead>
<tr>
<th></th>
<th>Housing dataset</th>
<th>Crime dataset</th>
<th>Loans dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>I</td>
<td>M</td>
</tr>
<tr>
<td>SRSWOR</td>
<td>3</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Stratified Sampling</td>
<td>7</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Combination</td>
<td>8</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Table: C = correct output, I = incorrect output, M = missing output.
London Crimes Data. Insight found grouping by:

- borough

- major crime category (high-level categorization of crime)

- minor crime category (low-level categorization of crime)

London Crimes Data. Insight found grouping by district.
If the Outstanding #1 condition is satisfied, this does not necessarily mean the result is interesting. Consider the output example from the Loans dataset:

<table>
<thead>
<tr>
<th>Application type</th>
<th>Loan request amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Application</td>
<td>$19,634.4</td>
</tr>
<tr>
<td>Individual</td>
<td>$14,788.28</td>
</tr>
</tbody>
</table>

The above is an easily inferable insight

Future work:
- Incorporate the ‘WHERE’ clause in an SQL query when searching for the Outstanding #1 insights.
- Estimate the sum of the measure after grouping.