Congressional samples
Juho Lamminmäki

Based on Congressional Samples for Approximate Answering of Group-By Queries (2000) by Swarup Acharyua et al.
Data Sampling

- Trying to obtain a maximally representative subset of the original data to reduce computation time or required storage.
- 100% accurate data is not always needed for analytics.
- The sample should work well with different kinds of queries.
Data Sampling

- The problem with plain uniform sampling
- Congressional samples
- Querying the sampled data
- Drawbacks of the approach
- Conclusion
Aggregation queries

SELECT sex, municipality, party, AVG(age)
FROM poll
WHERE election_year = 2017
GROUP BY sex, municipality, party
Uniform sampling

- Given a sample size \( X \) and the size of the original data \( D \), pick \( X \) random rows with an equal probability.
- However, if some groups are very small, only a few rows are picked from those groups.
- Accuracy becomes an issue with very small samples.
The basic idea behind the solution

- A larger proportion of the original group has to be sampled if the group is small.
- Fewer rows can be sampled from the larger groups since the accuracy does not suffer as much.
- Uniform sampling is important because it works the best if the sample is later queried using predicates.
Congressional samples

- House
- Senate
- Basic Congress
- Congress
House

- Uniform sampling over the whole data.
House

Original

House

H1  H2  H3
Senate

- Given $m$ groups and a sample size $X$, take a sample of $X/m$ rows from each group, i.e. the total sample size is divided equally between all groups.
- May use too few samples from the larger groups.
House and Senate
Basic Congress

- A combination of House and Senate
- For each group $g$, the sample size is $\max(H_g, S_g)$ where $H_g$ and $S_g$ are the expected sample sizes of group $g$ in House and Senate sampling methods respectively.
House, Senate and Basic Congress

Original

House

H1  H2  H3

Senate

S1  S2  S3

Basic Congress (before scaling)

H1  S2  S3
Basic Congress

- Produces a total sample size $\geq X$, so the sample sizes of each group have to be scaled with a constant so that the total sample size becomes $X$. 
House, Senate and Basic Congress

Original

House

Senate

Basic Congress (before scaling)

Basic Congress (after scaling)
Not perfect
• Let A and B be some grouping attributes that group the data into four groups i.e. **GROUP-BY A, B**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>avg(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a1</td>
<td>b1</td>
<td>...</td>
</tr>
<tr>
<td>a1</td>
<td>b2</td>
<td>...</td>
</tr>
<tr>
<td>a1</td>
<td>b3</td>
<td>...</td>
</tr>
<tr>
<td>a2</td>
<td>b3</td>
<td>...</td>
</tr>
</tbody>
</table>
### Basic Congress

**Grouping attributes: A**

<table>
<thead>
<tr>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td>25%</td>
</tr>
</tbody>
</table>

**Grouping attributes: A, B**

<table>
<thead>
<tr>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
</tr>
</tbody>
</table>
Basic Congress

Grouping attributes: A

<table>
<thead>
<tr>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td>25%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>40%</td>
</tr>
</tbody>
</table>

As a percentage of the total sample size

Grouping attributes: A, B

<table>
<thead>
<tr>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27%</td>
<td>27%</td>
<td>23%</td>
<td>23%</td>
</tr>
</tbody>
</table>

As a percentage of the total sample size
## Basic Congress

**Grouping attributes: A**

<table>
<thead>
<tr>
<th></th>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td>25%</td>
<td></td>
</tr>
</tbody>
</table>

**Grouping attributes: A, B**

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
<td></td>
</tr>
</tbody>
</table>

**Grouping attributes: A**

<table>
<thead>
<tr>
<th></th>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>40%</td>
<td></td>
</tr>
</tbody>
</table>

**Grouping attributes: A, B**

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>23%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27%</td>
<td>27%</td>
<td>23%</td>
</tr>
</tbody>
</table>
# Basic Congress

## Grouping attributes: A

<table>
<thead>
<tr>
<th></th>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td>25%</td>
<td></td>
</tr>
</tbody>
</table>

**Optimal**

<table>
<thead>
<tr>
<th></th>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>40%</td>
<td></td>
</tr>
</tbody>
</table>

## Grouping attributes: A, B

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
<td></td>
</tr>
</tbody>
</table>

**Not optimal**

<table>
<thead>
<tr>
<th></th>
<th>(a1)</th>
<th>(a2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>23%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27%</td>
<td>27%</td>
<td>23%</td>
<td>23%</td>
<td></td>
</tr>
</tbody>
</table>

**Optimal**

Juho Lamminmäki
Congress

- A solution to the problem i.e. it works better than Basic Congress with subsets of the original grouping attributes.
- An extension of the basic congress
Congress

- All subsets of the grouping attributes are $\emptyset$, \{A\}, \{B\} and \{A, B\}.
- First, calculate the amount of groups created by each subset.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Groups</th>
<th>Total #</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>The whole data</td>
<td>1</td>
</tr>
<tr>
<td>{A}</td>
<td>(a1), (a2)</td>
<td>2</td>
</tr>
<tr>
<td>{B}</td>
<td>(b1), (b2), (b3)</td>
<td>3</td>
</tr>
<tr>
<td>{A, B}</td>
<td>(a1, b1), (a1, b2), (a1, b3), (a2, b3)</td>
<td>4</td>
</tr>
</tbody>
</table>
Congress

- Then, calculate the expected sample size for each group using senate sampling.
- If $X$ is the total sample size, then each group has a sample size of $X/(\text{number of groups})$.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Groups</th>
<th>Total #</th>
<th>Sample size of a single group</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>The whole data</td>
<td>1</td>
<td>X/1</td>
</tr>
<tr>
<td>{A}</td>
<td>(a1), (a2)</td>
<td>2</td>
<td>X/2</td>
</tr>
<tr>
<td>{B}</td>
<td>(b1), (b2), (b3)</td>
<td>3</td>
<td>X/3</td>
</tr>
<tr>
<td>{A, B}</td>
<td>(a1, b1), (a1, b2), (a1, b3), (a2, b3)</td>
<td>4</td>
<td>X/4</td>
</tr>
</tbody>
</table>
Congress

- So the expected sample size as a percentage of the total sample size $X$ for each group $(a_1, b_1), (a_1, b_2), (a_1, b_3), (a_2, b_3)$ becomes

<table>
<thead>
<tr>
<th></th>
<th>$(a_1, b_1)$</th>
<th>$(a_1, b_2)$</th>
<th>$(a_1, b_3)$</th>
<th>$(a_2, b_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
</tr>
<tr>
<td>${A}$</td>
<td>20%</td>
<td>20%</td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>${B}$</td>
<td>25%</td>
<td>25%</td>
<td>18.75%</td>
<td>31.25%</td>
</tr>
<tr>
<td>${A, B}$</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>
Congress

- The empty set does not group at all, so taking a senate sample with no grouping attributes is the same as taking a House (uniform) sample.

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>∅</td>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
</tr>
<tr>
<td>{A}</td>
<td>20%</td>
<td>20%</td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>{B}</td>
<td>25%</td>
<td>25%</td>
<td>18.75%</td>
<td>31.25%</td>
</tr>
<tr>
<td>{A, B}</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>
Congress

- Taking the maximum sample size from either $\emptyset$ or \{A, B\} and scaling is the same as Basic Congress

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
</tr>
<tr>
<td>{A, B}</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>MAX</td>
<td>30%</td>
<td>30%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>
Congress

• Adding the other subsets makes the Basic Congress into Congress.

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>∅</td>
<td>30%</td>
<td>30%</td>
<td>15%</td>
<td>25%</td>
</tr>
<tr>
<td>{A}</td>
<td>20%</td>
<td>20%</td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>{B}</td>
<td>25%</td>
<td>25%</td>
<td>18.75%</td>
<td>31.25%</td>
</tr>
<tr>
<td>{A, B}</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>MAX</td>
<td>30%</td>
<td>30%</td>
<td>25%</td>
<td>50%</td>
</tr>
</tbody>
</table>
Congress

- This ensures that the sample works reasonably well with any subset of the original grouping attributes.

<table>
<thead>
<tr>
<th></th>
<th>(a1, b1)</th>
<th>(a1, b2)</th>
<th>(a1, b3)</th>
<th>(a2, b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAX</td>
<td>30%</td>
<td>30%</td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td>SCALED</td>
<td>22.22%</td>
<td>22.22%</td>
<td>18.52%</td>
<td>37.04%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(a1)</th>
<th>(a2)</th>
<th>(b1)</th>
<th>(b2)</th>
<th>(b3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62.96%</td>
<td>37.04%</td>
<td>22.22%</td>
<td>22.22%</td>
<td>55.56%</td>
</tr>
</tbody>
</table>
Querying sampled data

- Averages, medians etc. work fine without modifications.
- Sums, counts etc. require modification.
Querying sampled data

- SELECT $\text{sum(value) \ast \frac{original\_size}{sample\_size}}$
  - Works only for uniform samples since $\frac{original\_size}{sample\_size}$ is not the correct “scale factor” for all groups in non-uniform (biased) samples.

- Storing the scale factor for each row
  - Very high maintenance overhead.

- Storing the scale factor for each group
  - Most likely the best solution
Querying sampled data

```
FROM values v
JOIN scale_factors s USING(A, B, C)
```

- Can be optimized further, but this is the basic idea.
- The scale factors have to be constantly maintained, but the overhead is not very high.
Drawbacks

- For some data, uniform sampling over the whole data, which is much easier to implement and maintain, may be good enough.
- Such data might be something where not many grouping attributes are needed and/or there exists no small groups.
Drawbacks

- Senate sampling (used in Congress and Basic Congress too) might try to sample more rows than there are in the original data.
- The original paper simply states that handling these scenarios is not straightforward and leaves it at that.
Drawbacks

- Aggregate attributes with a very high variance or outliers with extreme values do not behave well when uniformly sampled.
- e.g. \( \text{avg}(-3, 0, 3, 1, 1, 100000) = 16667 \), but \( \text{avg}(-3, 0, 3, 1) = 0.5 \)
Drawbacks

• In these cases, implementing a solution that buckets the values into ranges \([v_1, v_n] = [v_1, v_2] \cup \ldots \cup [v_{n-1}, v_n]\) and takes a representative sample from each bucket will yield better results (Error-bounded Sampling for Analytics on Big Sparse Data, Yin Yan et al., 2014).

• This kind of a solution is more accurate in general, but it is less flexible with e.g. query predicates and the aggregate attributes must be known beforehand.
Conclusion

- Data sampling is useful when saving resources or time trumps accuracy.
- Small groups a problem with uniform sampling.
- Congress sampling fixes the problem with small groups, but does not handle situations where the aggregate attribute has some extreme values.
- Sampling makes querying more complex.
Phew, it’s finally over!

In case you missed it, my name is Juho Lamminmäki