Panelists

- Bill Dimm, Hot Neuron
- Lilith Bat-Leah, Fronteo
- Cynthia Vasquez, CHAT Consulting
- Grady Glover, The Rodarti Group & Lawrence Bartels
- Tammi Loveland, U.S. DoJ
Prevalence

- Percentage having some property (relevant)
Recall

- Percentage of relevant docs found
- Defensibility
Precision

- Percentage of retrieved docs that are relevant
- Cost (review effort)
Sample 2 Docs

\[ n = 2 \]

Percent Relevant in Sample

0 20 40 60 80 100
Sample 10 Docs

$n = 2$

$n = 10$

Percent Relevant in Sample

0  20  40  60  80  100
Sample 30 Docs

$n = 10$

$n = 30$

Percent Relevant in Sample
Anything Is Possible
What Is Plausible?
Example of Finding a Confidence Interval

- Sample 25 documents, 18 are relevant
  - 72% of sample is relevant
- What is reasonable prevalence for population?
  - Point estimate is 72%
Plausible Population Prevalence? NO

\[ p = 10\% \]
Plausible Population Prevalence? NO

\[ p = 20\% \]

Percent Relevant in Sample
Plausible Population Prevalence? NO

$p = 30\%$

Percent Relevant in Sample
Plausible Population Prevalence? YES

$p = 50.7\%$
Plausible Population Prevalence? YES

$p = 50.7\%$

Percent Relevant in Sample
Plausible Population Prevalence? YES

\[ p = \hat{p} \approx 60\% \]
Plausible Population Prevalence? YES

\[ p = 50.7\% \quad p = 70\% \]
Plausible Population Prevalence? YES

\[ p = 50.7\% \]

\[ p = 80\% \]
Plausible Population Prevalence? YES

\[ p = 50.7\% \]

\[ p = 87.9\% \]
Plausible Population Prevalence? YES

\[ p = 50.7\% \]

\[ p = 87.9\% \]
Plausible Population Prevalence? NO

\[ p = 50.7\% \]

\[ p = 87.9\% \]

\[ p = 90\% \]
Confidence Interval

\[ p = 50.7\% \]

\[ p = 87.9\% \]
Wald Approximation

\[
\delta p = 1.96 \sqrt{\frac{p(1-p)}{S}}
\]

\[
\delta p_{\text{max}} = \frac{0.98}{\sqrt{S}}
\]

\[
S = \frac{0.96}{\delta p_{\text{max}}^2}
\]
Standard Sample Sizes, 95% Confidence

<table>
<thead>
<tr>
<th>$\delta \rho_{\text{max}}$</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>9,600</td>
</tr>
<tr>
<td>2%</td>
<td>2,400</td>
</tr>
<tr>
<td>3%</td>
<td>1,067</td>
</tr>
<tr>
<td>4%</td>
<td>600</td>
</tr>
<tr>
<td>5%</td>
<td>384</td>
</tr>
<tr>
<td>7%</td>
<td>196</td>
</tr>
<tr>
<td>10%</td>
<td>96</td>
</tr>
<tr>
<td>15%</td>
<td>43</td>
</tr>
</tbody>
</table>
Example

- Sample 10,000 Docs From Full Population
  - 100 are relevant
  - 200 match search query
  - 80 are relevant and match search query

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Point Est</th>
<th>Sample Size</th>
<th>Worst Case CI</th>
<th>Wald CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence</td>
<td>1%</td>
<td>10,000</td>
<td>±1%</td>
<td>±0.2%</td>
</tr>
<tr>
<td>Precision</td>
<td>40%</td>
<td>200</td>
<td>±7%</td>
<td>±6.8%</td>
</tr>
<tr>
<td>Recall</td>
<td>80%</td>
<td>100</td>
<td>±10%</td>
<td>±7.8%</td>
</tr>
</tbody>
</table>
Elusion Sampling

- **Synonyms:**
  - Null Set
  - Elusion Set
  - Discard Set
  - Negatives

- **Problems**
  - Bias (no fix)
  - Sample Size
Recall From Elusion

\[ R \]

\[ E \]
Elusion Sampling

\[ R = \frac{TP}{TP + (N - n)E} \]

\[ \delta R = 1.96 \sqrt{\frac{E(1 - E)}{S}} \frac{(N - n)TP}{[TP + (N - n)E]^2} \]

\[ \delta R_{\text{max}} = \frac{0.98}{\sqrt{S}} \frac{9}{8\sqrt{3}} \sqrt{\frac{N - n}{TP}} \]

\[ S = \frac{0.96}{\delta R_{\text{max}}^2} \frac{27}{64} \frac{N - n}{TP} \]
Elusion vs. Direct Method

<table>
<thead>
<tr>
<th>$R$</th>
<th>$S_{\text{Elusion}}/S_{\text{Direct}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0.47</td>
</tr>
<tr>
<td>0.80</td>
<td>0.53</td>
</tr>
<tr>
<td>0.70</td>
<td>0.60</td>
</tr>
<tr>
<td>0.60</td>
<td>0.70</td>
</tr>
<tr>
<td>0.50</td>
<td>0.84</td>
</tr>
<tr>
<td>0.40</td>
<td>1.05</td>
</tr>
<tr>
<td>0.30</td>
<td>1.41</td>
</tr>
<tr>
<td>0.20</td>
<td>2.11</td>
</tr>
<tr>
<td>0.10</td>
<td>4.22</td>
</tr>
</tbody>
</table>
Lies, Damn Lies, and Statistics

- Inclusive Emails
- Search Terms