We have done that already…

- We built a model for how users “surf” the web
  - Allowed us to answer questions about relative popularity of different web pages
- We built a model for how queues evolve
  - Allowed us to answer questions about implications on queuing delay from increased load

But, are these models any good?

- We already know that we made some possibly unwarranted assumptions — e.g.,
  - People click on links “randomly”
  - Queues can hold as many packets as necessary
- We need to “validate” our models/assumptions
  - We validate a model (or assumption) by measuring the real artifact and comparing the results to what the model predicts (or what the assumption states)

Let’s try an exercise…

- Assumption: Number of hops between two computers is a good indicator of (is correlated with) the propagation delay between them
- Why is this a useful assumption?
  - Counting hops is easier than measuring propagation delays
  - Instrumental for server selection purposes (e.g., Akamai)

How do we validate this statement?

- Measure the number of hops and the propagation delay between every pair of Internet computers to find out if relationship is evident
- How many pairs of computers are there?
  - Impossible!
Statistics to the rescue!

- Measure the number of hops and the propagation delay between sampled pairs of Internet computers to find out if relationship is evident

- How do we measure?

- How do we sample?

- How do we establish relationship?

How do we measure? Traceroute

![Traceroute Image]

How do we sample?

- Simple Random Sampling (SRS)
  - Hard! Cannot run traceroute from arbitrary computers...

- Traceroute is installed on a number of computers around the world
  - List available at traceroute.org

- Traceroute "vantage points" may not be representative
  - Use stratified sampling

Let's try one stratum: ".edu" in US

- Vantage Points:
  - CMU
  - BU
  - MIT
  - USC
  - WISC
  - Washington
  - Arizona
  - San Diego
  - Stanford
  - Berkeley

63 "successful" experiments

![Graph Image]

How do we establish “relationship”? 

- Recall:
  
  An association exists between two variables if a particular value of one variable is more likely to occur with certain values of the other variable.

- Two variables
  - Response Variable: Propagation Delay
  - Explanatory Variable: Number of hops
Using hops to explain delays

- Number of paths with
  - Small vs large hop-count
  - Small vs large propagation delay

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>Delay &lt; 50ms</th>
<th>Delay &gt;= 50ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hops &lt; 12</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Hops &gt;= 12</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Results: Hops vs delay

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Delay &lt; 50ms</th>
<th>Delay &gt;= 50ms</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hops &lt; 12</td>
<td>22</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>Hops &gt;= 12</td>
<td>9</td>
<td>28</td>
<td>37</td>
</tr>
</tbody>
</table>

What are the odds?

- Sample Odds Ratio = n00*n11/n01*n10
- $\hat{OR} = \frac{22*28}{9*4} = 17.11$
- $\hat{OR} = 17.11$
- $\ln(\hat{OR}) = 2.84$
- $\text{SD} = \sqrt{\frac{1}{22}+\frac{1}{28}+\frac{1}{9}+\frac{1}{4}} = 0.665$

95% Confidence Interval for $\ln(\hat{OR})$ is given by $(\ln(\hat{OR}) - 2*SD, \ln(\hat{OR}) + 2*SD)$

- 95% CI for $\ln(\hat{OR})$ = (1.51, 4.17)
- 95% CI for OR around 17.11 = (4.52, 64.7)

Are all hops significant w.r.t. delay?

- Local hops seem to contribute insignificant delay
- Perhaps we get better association if we only consider "backbone" hops
Results: Backbone hops vs delay

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Delay &lt; 50ms</th>
<th>Delay &gt;= 50ms</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHops &lt; 7</td>
<td>30</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>BHops &gt;= 7</td>
<td>1</td>
<td>27</td>
<td>28</td>
</tr>
</tbody>
</table>

What are the odds?

- OR = 162
- ln(OR) = 5.09
- SD = 1.13
- 95% CI for ln(OR) = (2.83, 7.35)
- 95% CI for OR around 162 = (16.95, 1556.2)
- Can confidently assume association!

Results: Local hops vs delay

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Delay &lt; 50ms</th>
<th>Delay &gt;= 50ms</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>LHops &lt; 6</td>
<td>8</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>LHops &gt;= 6</td>
<td>23</td>
<td>26</td>
<td>49</td>
</tr>
</tbody>
</table>

What are the odds?

- OR = 1.5
- ln(OR) = 0.405
- SD = 0.611
- 95% CI for ln(OR) = (-0.817, 1.627)
- 95% CI for OR around 1.5 = (0.442, 5.09)
- Cannot confidently assume association!
Let’s try another exercise…

- Assumption: Physical distance between two computers is a good indicator of (is correlated with) the propagation delay between them.

Results: Distance vs Delay

<table>
<thead>
<tr>
<th>Distance in miles</th>
<th>Delay &lt; 50ms</th>
<th>Delay &gt;= 50ms</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>D &lt; 1.5K miles</td>
<td>30</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>D &gt;= 1.5K miles</td>
<td>0</td>
<td>32</td>
<td>32</td>
</tr>
</tbody>
</table>

Results: Hops vs delay

- Other approaches (beyond MCS-109) exist for checking associations.
- Example: Fisher’s exact test:
  - Quantifies the probability that the observed results are due to pure chance.
  - Recall the “monkey and the keyboard.”
Results: Backbone hops vs delay

Fisher's Exact Test

<table>
<thead>
<tr>
<th>TABLE</th>
<th>50, 5, 1, 27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>p-value = 0.9999999999999428</td>
</tr>
<tr>
<td>Right</td>
<td>p-value = 9.977011429412016e-12</td>
</tr>
<tr>
<td>2-Tail</td>
<td>p-value = 1.1584133115850835e-11</td>
</tr>
</tbody>
</table>

Results: Local hops vs delay

Fisher's Exact Test

<table>
<thead>
<tr>
<th>TABLE</th>
<th>8, 6, 23, 26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>p-value = 0.835465505091507</td>
</tr>
<tr>
<td>Right</td>
<td>p-value = 0.35574129679271127</td>
</tr>
<tr>
<td>2-Tail</td>
<td>p-value = 0.5561270670544927</td>
</tr>
</tbody>
</table>

Results: Distance vs delay

Fisher's Exact Test

<table>
<thead>
<tr>
<th>TABLE</th>
<th>30, 1, 0, 32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>p-value = 1</td>
</tr>
<tr>
<td>Right</td>
<td>p-value = 3.6013931348768953e-17</td>
</tr>
<tr>
<td>2-Tail</td>
<td>p-value = 3.6013931348768953e-17</td>
</tr>
</tbody>
</table>

What kind of association?

Total Hops vs Delay

R^2 = 0.5856

Percentage of variation in delay "explained by" the number of hops.

What kind of association?

Backbone Hops vs Delay

R^2 = 0.7276
What kind of association?

Distance vs Delay

$R^2 = 0.9333$

70.00
80.00
90.00
100.00

MA/CS-109 (Azer Bestavros) 4011/29/2010

Footnotes and desiderata

- Not quite SRS
  - Not all .edu computers were equally likely to be selected as vantage points for traceroute – 6 in PST/MST time zones and 3 in EST/CST time zones

- Non-response bias
  - Traceroute did not succeed with some of the target .edu computers

- What are the “right” strata for sampling?