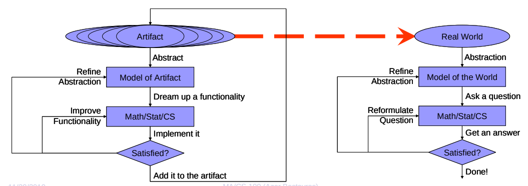
MA/CS-109: Delay Model Validation

Azer Bestavros

Understanding emergent behaviors

- Stacking up simple functionalities allowed us to build a complex artifact.
- Even though we “built it”, we often cannot explain some of its characteristics.
- Abstraction to the rescue again!



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)

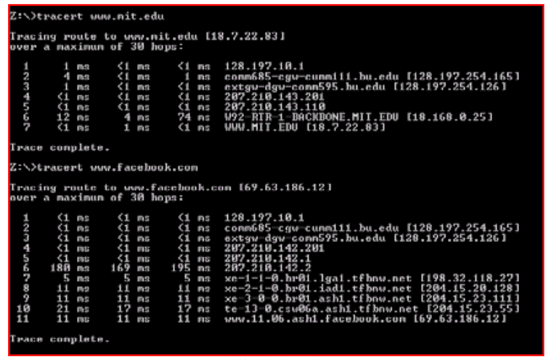
How do we validate this statement?

- 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



```
Z:\>tracert www.mit.edu

Tracing route to www.mit.edu [18.7.22.83]
over a maximum of 30 hops:

  1     1 ms    <1 ms    <1 ms  128.197.10.1
  2     4 ms    <1 ms     1 ms  comm685-cgw-cumm111.bu.edu [128.197.254.165]
  3     1 ms    <1 ms    <1 ms  extgw-dgw-comm595.bu.edu [128.197.254.126]
  4    <1 ms    <1 ms    <1 ms  207.210.143.201
  5    <1 ms    <1 ms    <1 ms  207.210.143.110
  6    12 ms     4 ms    74 ms  W92-RTR-1-BACKBONE.MIT.EDU [18.168.0.25]
  7    <1 ms     1 ms    <1 ms  WWW.MIT.EDU [18.7.22.83]

Trace complete.

Z:\>tracert www.facebook.com

Tracing route to www.facebook.com [69.63.186.12]
over a maximum of 30 hops:

  1    <1 ms    <1 ms    <1 ms  128.197.10.1
  2    <1 ms    <1 ms    <1 ms  comm685-cgw-cumm111.bu.edu [128.197.254.165]
  3    <1 ms    <1 ms    <1 ms  extgw-dgw-comm595.bu.edu [128.197.254.126]
  4    <1 ms    <1 ms    <1 ms  207.210.142.201
  5    <1 ms    <1 ms    <1 ms  207.210.142.1
  6   180 ms   169 ms   195 ms  207.210.142.2
  7     5 ms     5 ms     5 ms  xe-1-1-0.br01.lga1.tfbnw.net [198.32.118.27]
  8    11 ms    11 ms    11 ms  xe-2-1-0.br01.iad1.tfbnw.net [204.15.20.128]
  9    11 ms    11 ms    11 ms  xe-3-0-0.br01.ash1.tfbnw.net [204.15.23.111]
 10    21 ms    17 ms    17 ms  te-13-0.csw06a.ash1.tfbnw.net [204.15.23.55]
 11    11 ms    11 ms    11 ms  www.11.06.ash1.facebook.com [69.63.186.12]

Trace complete.
```

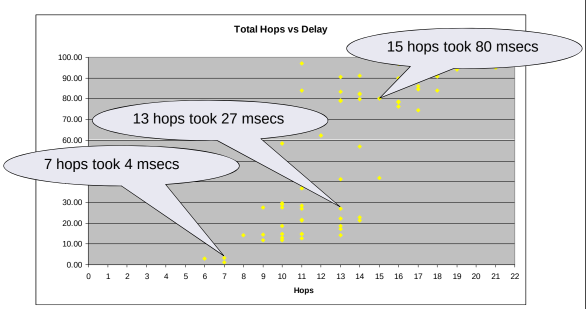

How do we sample?

- 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



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

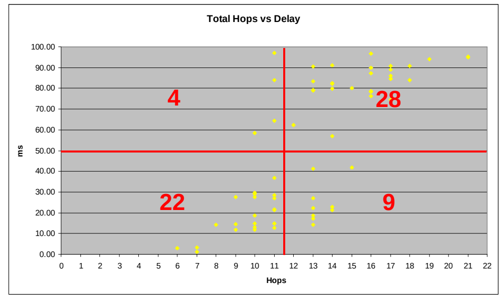
Using hops to explain delays

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

Explanatory Variable → rows; Response Variable → columns

	Delay < 50ms	Delay>= 50ms
Hops < 12	?	?
Hops >= 12	?	?

Results: Hops vs delay



Results: Hops vs delay

Frequency

	Delay < 50ms	Delay>= 50ms	All
Hops < 12	22	4	26
Hops >= 12	9	28	37

What are the odds?

	Delay < 50ms	Delay>= 50ms	All
Hops < 12	22	4	26
Hops >= 12	9	28	37

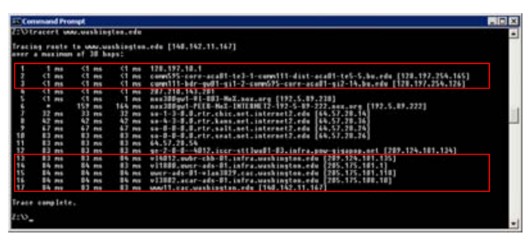
- Sample Odds Ratio = n00*n11/n01*n10
- $\widehat{OR}$ = 22*28 / 9*4 = 17.11
- **$\widehat{OR}$ = 17.11**
- **ln($\widehat{OR}$) = 2.84**
- Std Deviation = Sqrt(1/n00+1/n11+1/n01+1/n10)
- SD = Sqrt(1/22+1/28+1/9+1/4)
- **SD = 0.665**

What are the odds?

	Delay < 50ms	Delay>= 50ms	All
Hops < 12	22	4	26
Hops >= 12	9	28	37

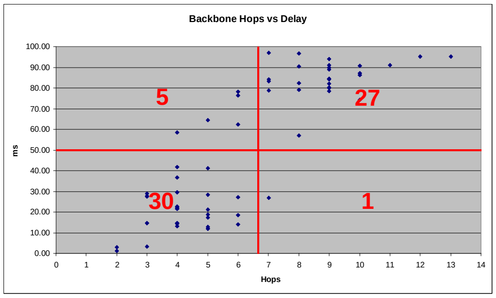
- 95% Confidence Interval for ln(OR) is given by (ln(OR) – 2*SD , ln(OR) + 2*SD)
- 95% CI for ln(OR) = (1.51,4.17)
- **95% CI for OR around 17.11 = (4.52,64.7)**

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



Results: Backbone hops vs delay

Frequency

	Delay < 50ms	Delay >= 50ms	All
BHops < 7	30	5	35
BHops >= 7	1	27	28

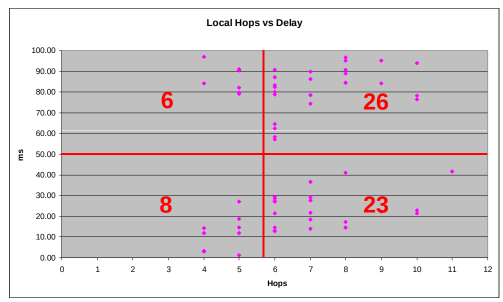
What are the odds?

Frequency

	Delay < 50ms	Delay >= 50ms	All
BHops < 7	30	5	35
BHops >= 7	1	27	28

- **$\widehat{OR}$ = 162**
- ln($\widehat{OR}$) = 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



Results: Local hops vs delay

Frequency

	Delay < 50ms	Delay >= 50ms	All
LHops < 6	8	6	14
LHops >= 6	23	26	49

What are the odds?

Frequency

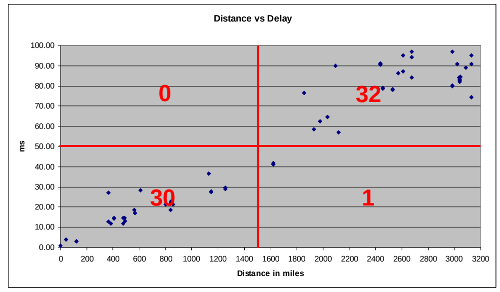
	Delay < 50ms	Delay >= 50ms	All
LHops < 6	8	6	14
LHops >= 6	23	26	49

- **$\widehat{OR}$ = 1.5**
- ln($\widehat{OR}$) = 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



Results: Distance vs delay

	Frequency		
	Delay < 50ms	Delay >= 50ms	All
D < 1.5K miles	30	1	31
D >= 1.5K miles	0	32	32

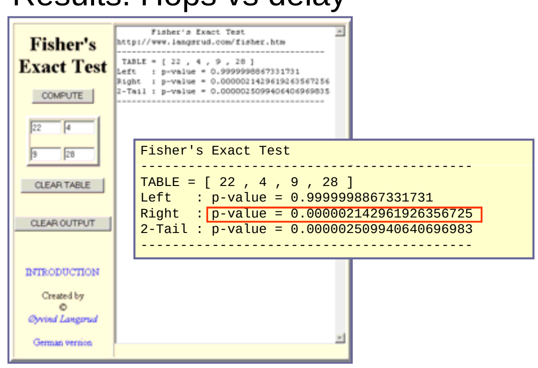
What are the odds?

- Approach we used in class cannot be used! Why?
  - Based on the CLT
  - CLT assumes a very large sample and that non-zero observations for all combinations of response and explanatory variables
- Important to remember assumptions and fine prints – remember: "garbage in garbage out"!

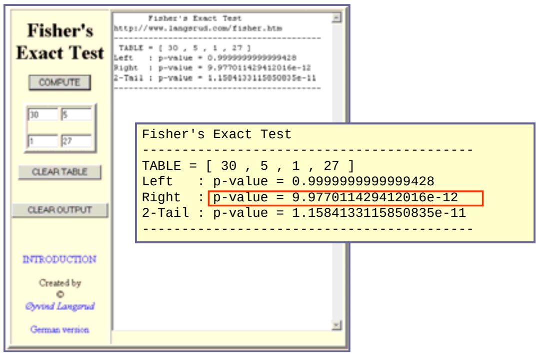
What are the odds?

- 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"

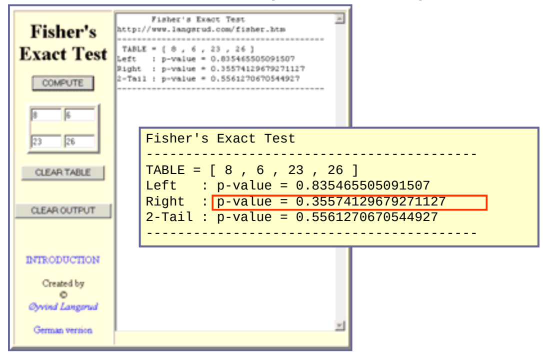
Results: Hops vs delay



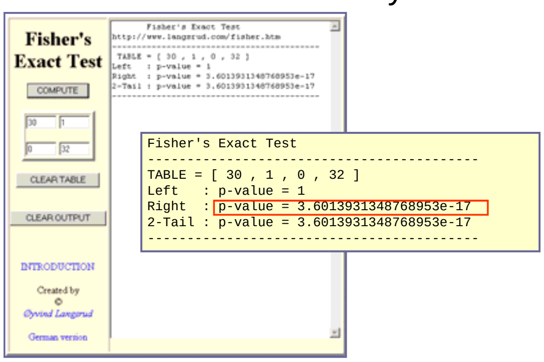
Results: Backbone hops vs delay



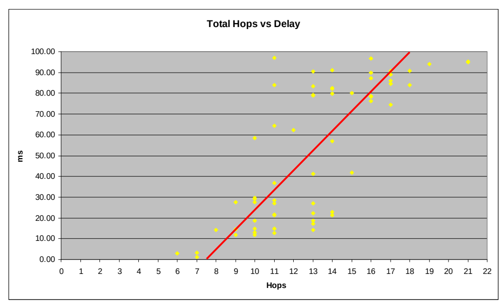
Results: Local hops vs delay



Results: Distance vs delay

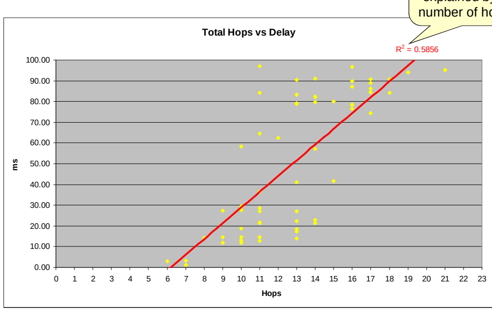


What kind of association?

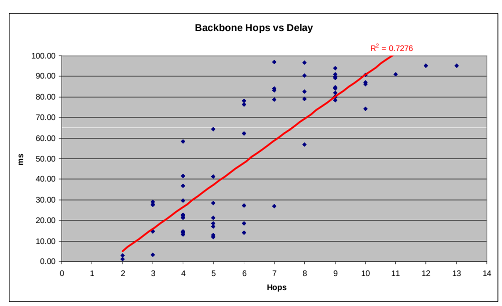


What kind of association?

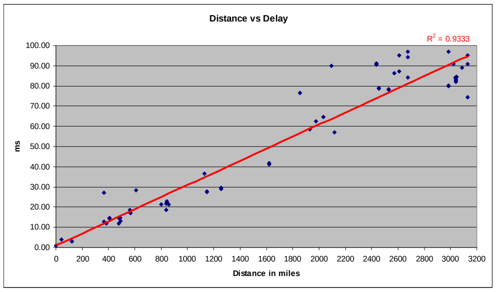
Percentage of variation in delay "explained by" the number of hops.



What kind of association?



What kind of association?



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?