Performance Evaluation: A Preparation for Statistics and Data Science ?

Jean-Yves Le Boudec

EPFL I&C, Lausanne, Switzerland

12 November 2021

T e a P A C S – Performance 2021

First International Workshop on Teaching Performance Analysis of Computer Systems

1. Probability As Seen By Students

Plausible, well accepted axioms about sample space and events

Frightening computations.

Sample space is not specified.

4.1 Axioms of probability

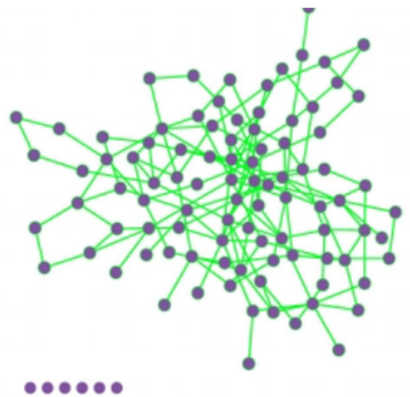
A **probability space** is a triple $(\Omega, \mathscr{F}, P)$, in which Ω is the sample space, $\mathscr{F}$ is a collection of subsets of Ω, and P is a **probability measure** $P : \mathscr{F} \to [0, 1]$.

[Weber]

Axioms of Probability:

- Axiom 1: For any event A, $P(A) \geq 0$.
- Axiom 2: Probability of the sample space S is $P(S) = 1$.
- Axiom 3: If $A_1, A_2, A_3, \cdots$ are disjoint events, then $P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$

[Pishro-Nik]

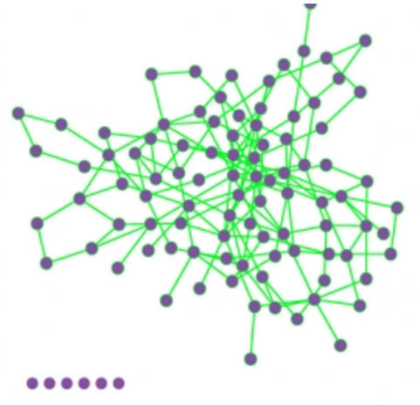


In a given realization of a random network some nodes gain numerous links, while others acquire only a few or no links (Image 3.3). These differences are captured by the degree distribution, p_k, which is the probability that a randomly chosen node has degree k. In this section we derive p_k for a random network and discuss its properties.

[Barabasi]

Network Science: Example of Student Assignment

Assignment: Consider a random graph with N vertices where every pair of vertices is connected with probability q. Compute the probability p_k that a vertex has degree k.



[Barabasi]

What is the sample space ?

What does probability p_k really mean ?

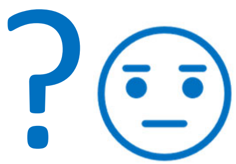
How do I compute p_k ?

Textbook Solution

Assignment: Consider a random graph with N vertices where every pair of vertices is connected with probability q. Compute the probability p_k that a vertex has degree k.

Answer:

$$p_k = \binom{N-1}{k} q^k (1-q)^{N-1-k}$$



"In a random network the probability that node i has exactly k links is the product of three terms:

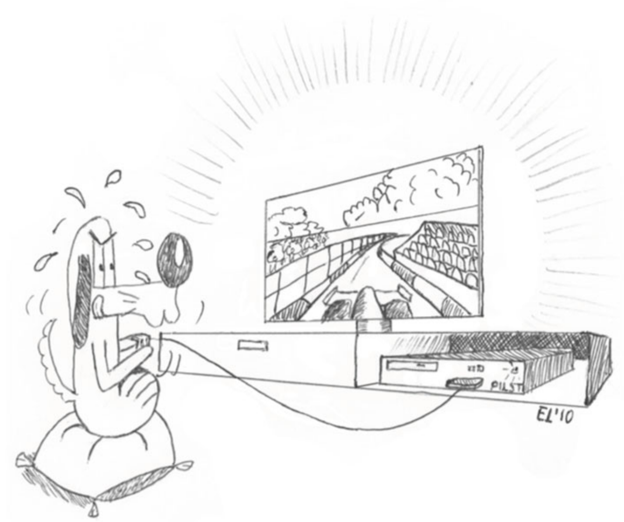
- *The probability that k of its links are present, or q^k.*
- *The probability that the remaining $(N-1-k)$ links are missing, or $(1-q)^{N-1-k}$*
- *The number of ways we can select k links from $N-1$ potential links a node can have, or $\binom{N-1}{k}$*

[Barabasi]

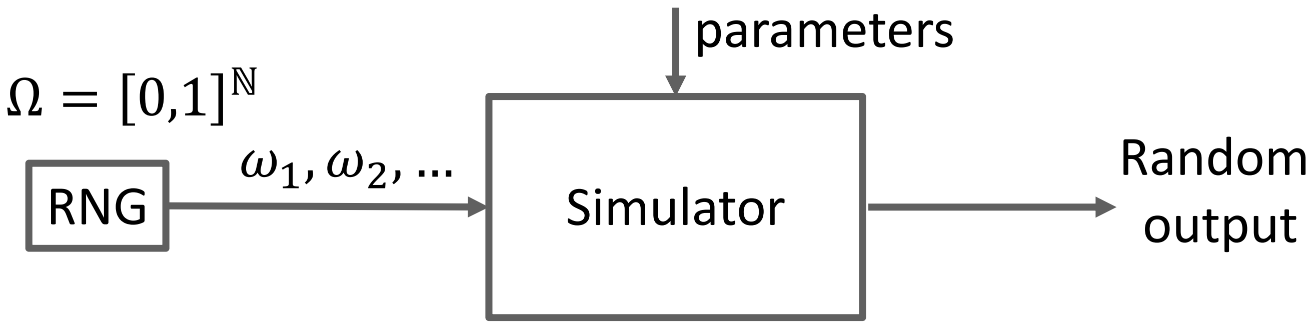
Performance Evaluation : Example of Student Assignment

Write a simulator.

A common and (enjoyable) exercise in a performance evaluation course.

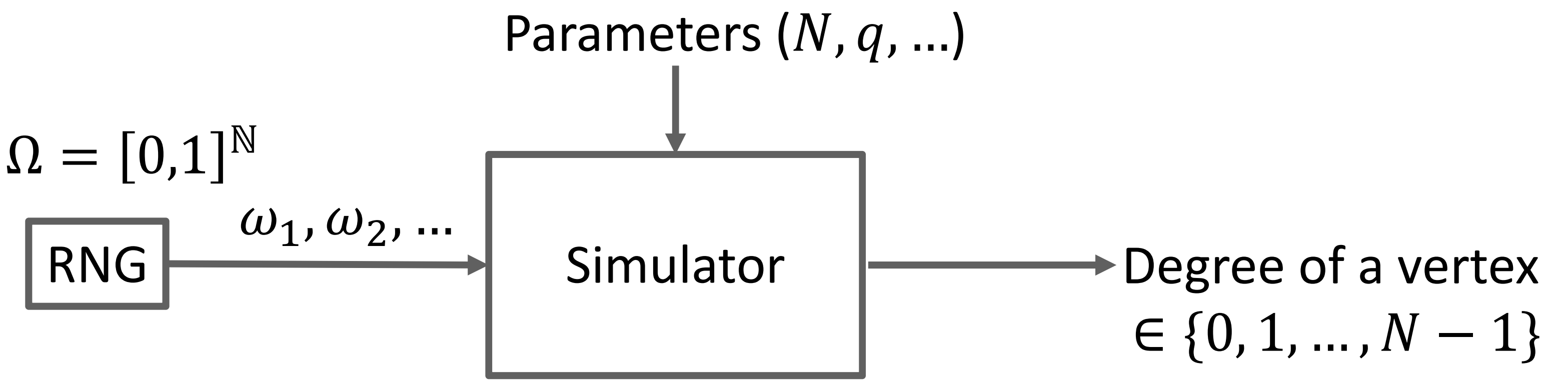


Well specified. Replaces mythical sample space by pseudo-random number generator. Probabilistic statements about output are well defined.



Solving the Network Science Assignment: The PE way

Assignment: Consider a random graph with N vertices where every pair of vertices is connected with probability q. Compute the probability p_k that a vertex has degree k.



We could run the simulator a very large of times and obtain the answer with a confidence interval.

But we can also use the simulator as a thought experiment to tackle the theoretical problem. Reveals that assignment needs more assumptions.

Interpretation 1

Assignment: Consider a random graph with N vertices where every pair of vertices is connected with probability q. Compute the probability p_k that a vertex has degree k.

N (number of vertices)
q (proba of edge)

RNG → For every (i, j) with $1 \leq i < j \leq N$, draw an edge with prob q. The draws are independent. → One random graph

Answer 1: $p_k = \frac{1}{N} \times$# vertices that have degree k

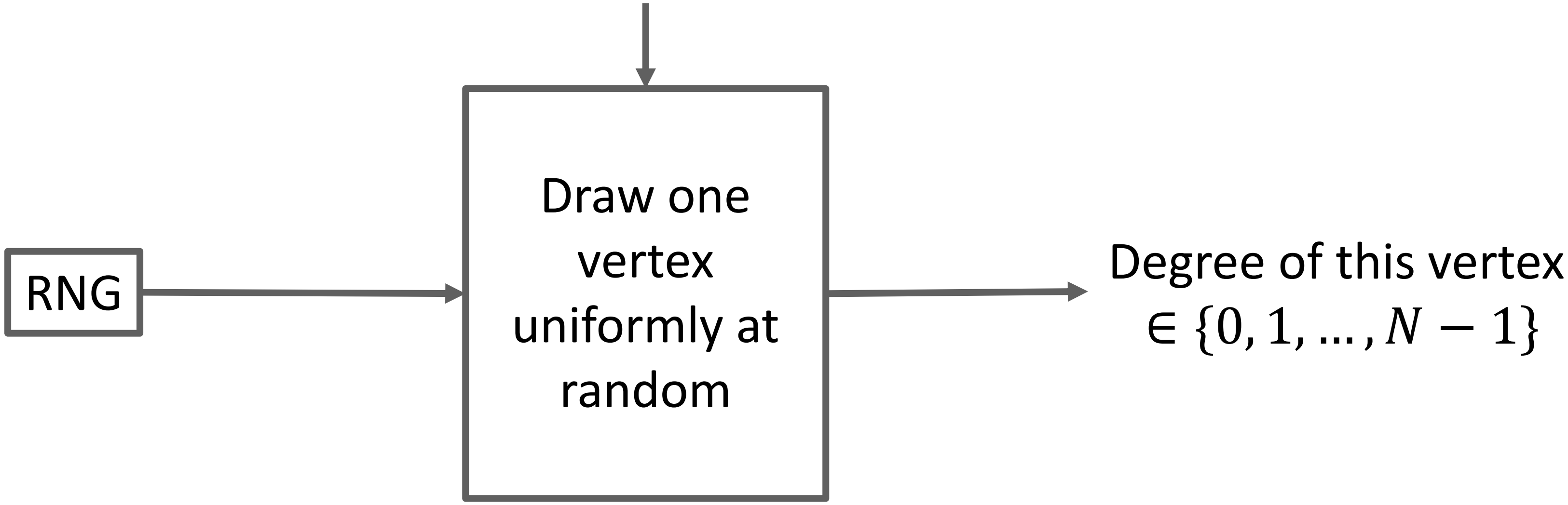
$(p_0, \dots, p_k, \dots, p_{N-1})$ is random !

Not the textbook anwer !

Interpretation 2

Assignment: Consider a random graph with N vertices where every pair of vertices is connected with probability q. Compute the probability p_k that a vertex has degree k.

One graph with N vertices
s.t. proportion of connected pairs is q



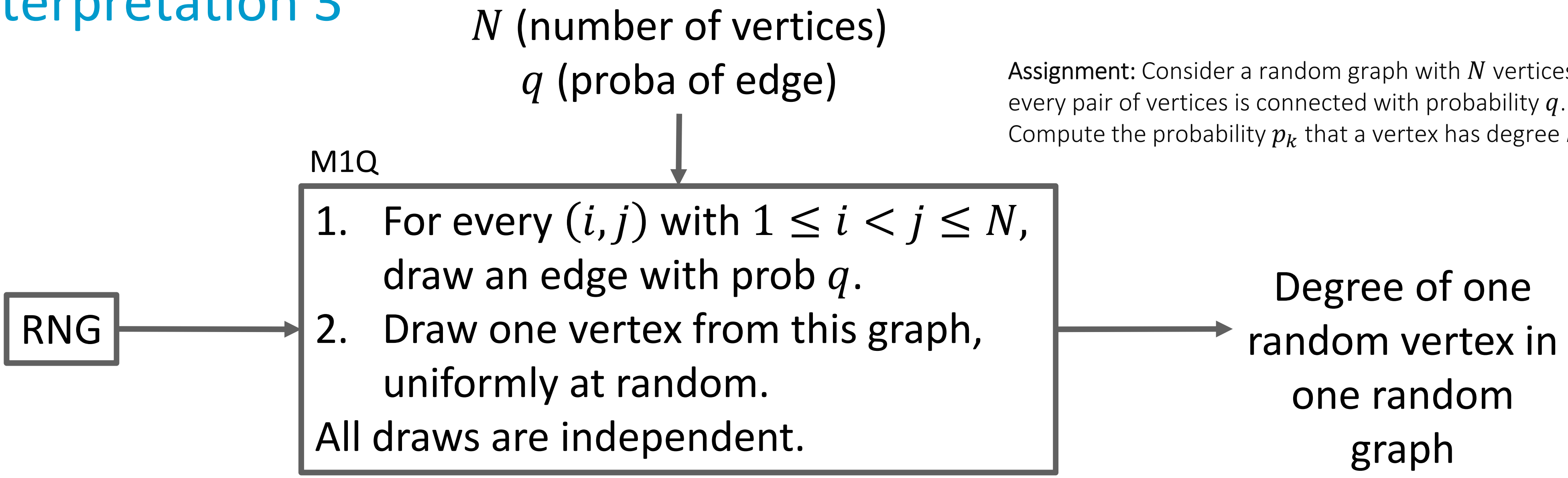
Answer 2: p_k = probability that the output is k

Depends on the input graph, not just N, q, k

Not the textbook anwer !

Interpretation 3

Assignment: Consider a random graph with N vertices where every pair of vertices is connected with probability q. Compute the probability p_k that a vertex has degree k.



Answer 3: $p_k = ?$

Seems compatible with the texbook answer.

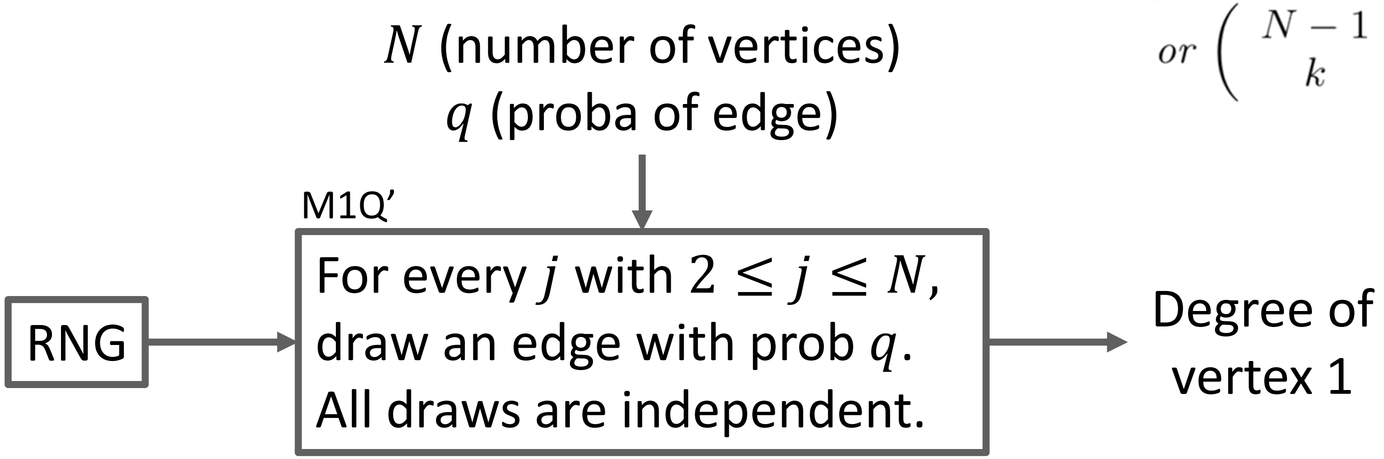
How do we compute it ? Let us look at the textbook answer.

Interpretation of the Textbook Answer

$$p_k = \binom{N-1}{k} q^k (1-q)^{N-1-k}$$

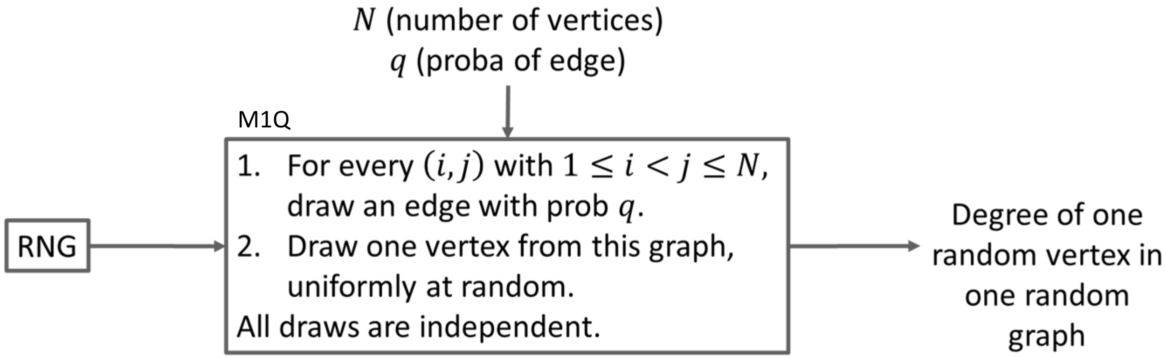
"In a random network the probability that node i has exactly k links is the product of three terms:

- *The probability that k of its links are present, or q^k.*
- *The probability that the remaining $(N-1-k)$ links are missing, or $(1-q)^{N-1-k}$*
- *The number of ways we can select k links from $N-1$ potential links a node can have, or $\binom{N-1}{k}$* [Barabasi]

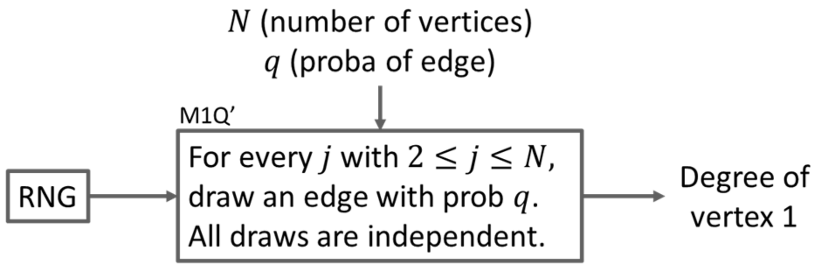


Not the same simulator as intepretation 3 !

An Understandable Proof of the Textbook Answer



Student's model



Model that supports the textbook answer

Proof: the outputs of the two simulators have same distribution

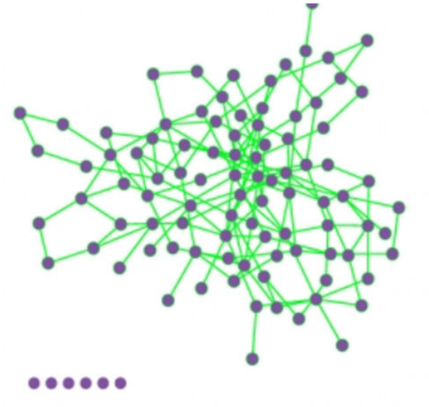
Student's model outputs $D = \sum_{1\leq i<I} X_{i,I} + \sum_{I<i\leq N} X_{I,i}$ where I ~Uniform in $\{1,\ldots,N\}$ and $X_{i,j}$ are Bernoulli (q), all independent.

Textbook model outputs $D' = \sum_{2\leq i\leq \mathrm{N}} Y_i$ where Y_i are i.i.d. Bernoulli (q).

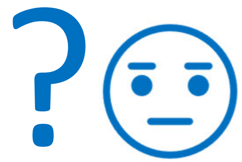
$$\mathbb{P}(D = k \mid I = i_0) = \mathbb{P}(D' = k)$$

$$\mathbb{P}(D = k) = \sum_i \mathbb{P}(D = k \mid I = i_0)\mathbb{P}(I = i_0) = \mathbb{P}(D' = k)\sum_{i_0} \mathbb{P}(I = i_0) = \mathbb{P}(D' = k).$$

"In a given realization of a random network some nodes gain numerous links, while others acquire only a few or no links (Image 3.3). These differences are captured by the degree distribution, p_k, which is the probability that a randomly chosen node has degree k". [Barabasi]



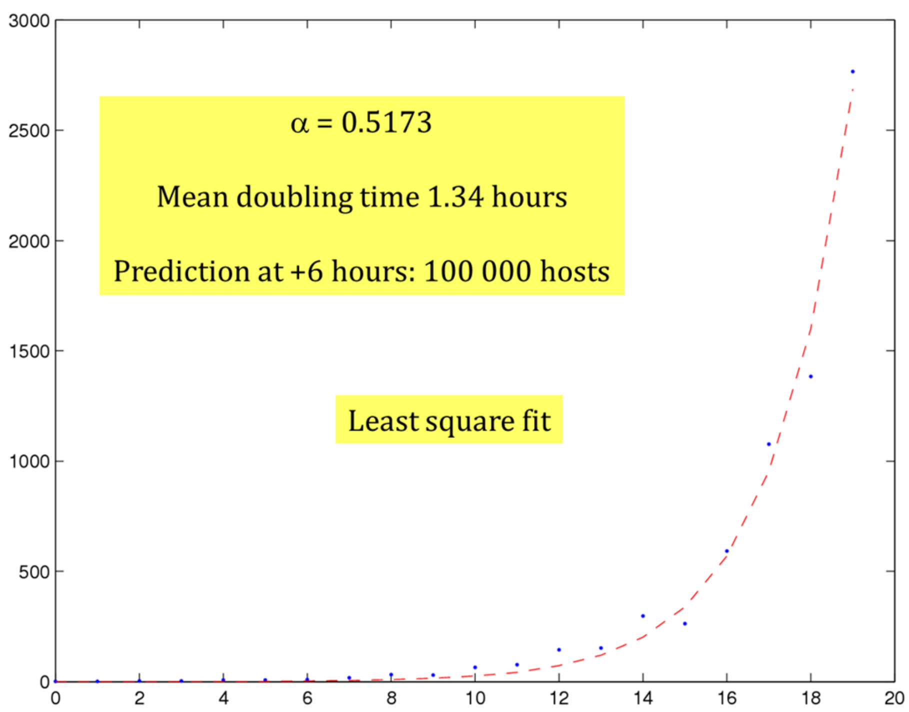
Tetxbook uses two inconsistent interpretations: the formula applies to interpretation 3, but is used with interpretation 2.

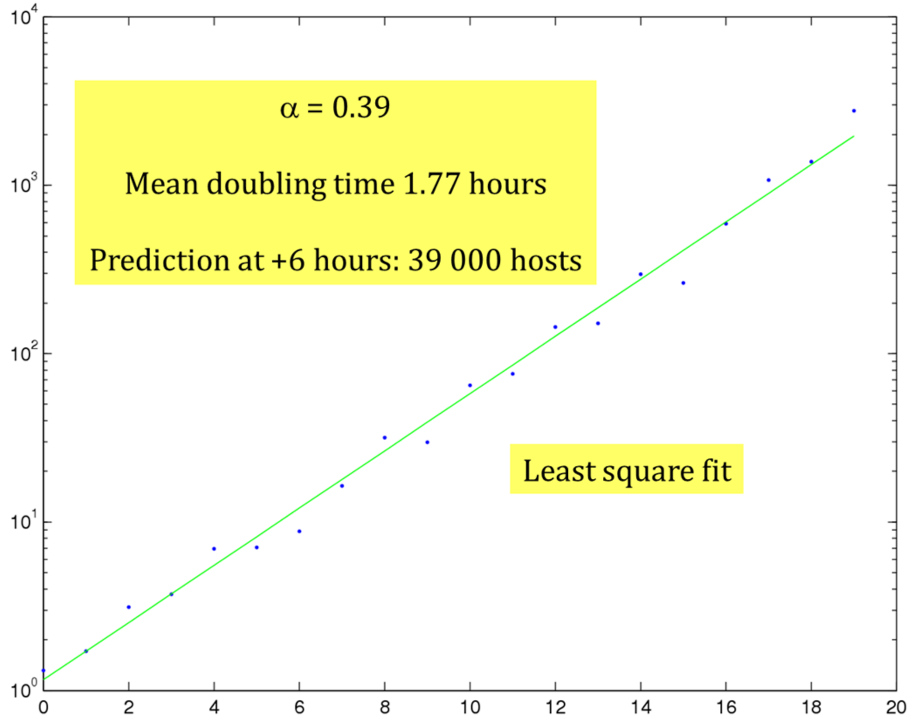


Take home message: Inconsistent interpretations of probabilities can be avoided by using simulators as thought experiments to specify the model.

2. Data Science Example

Estimate growth rate of viral infection



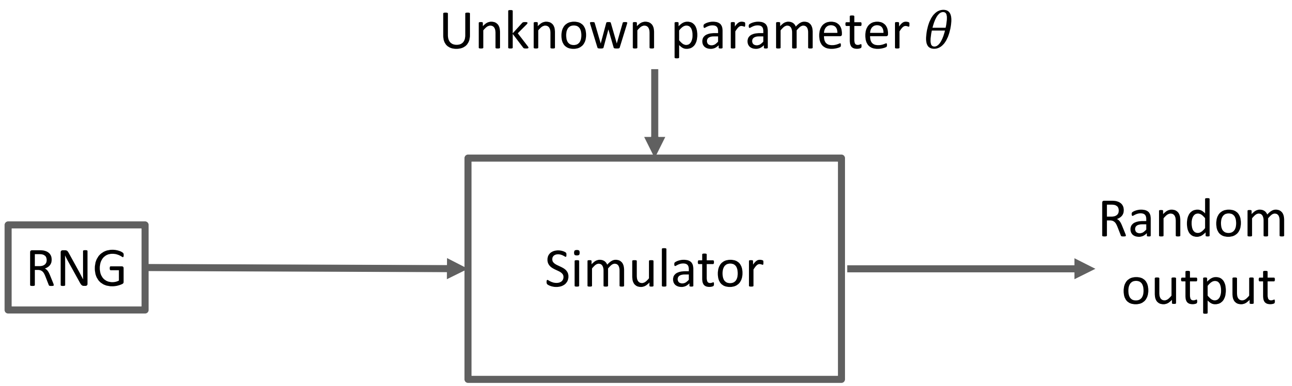


Classical Statistics Helps Determine Best Model

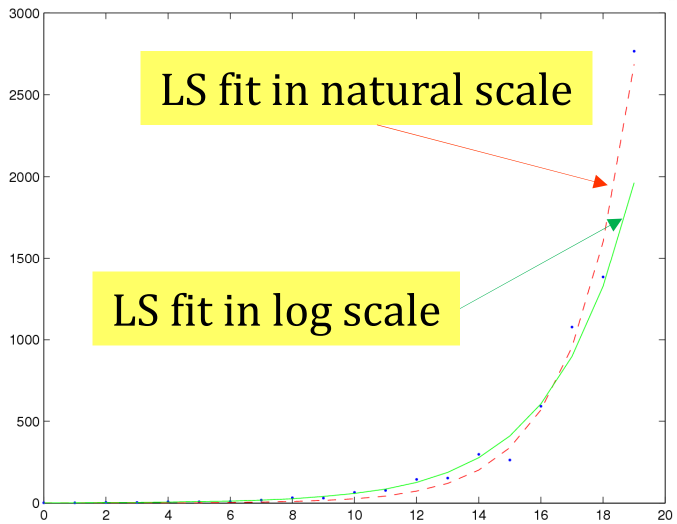
Classical statistics expressed in Performance Evaluation parlance :

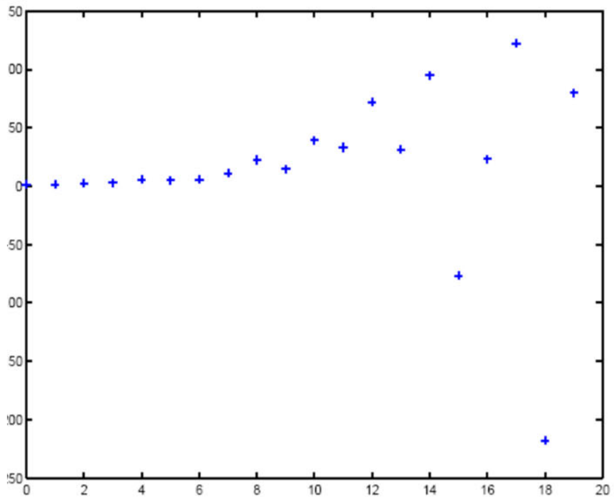
data at hand is one output of simulator with unknown parameter θ

Goal is to estimate θ

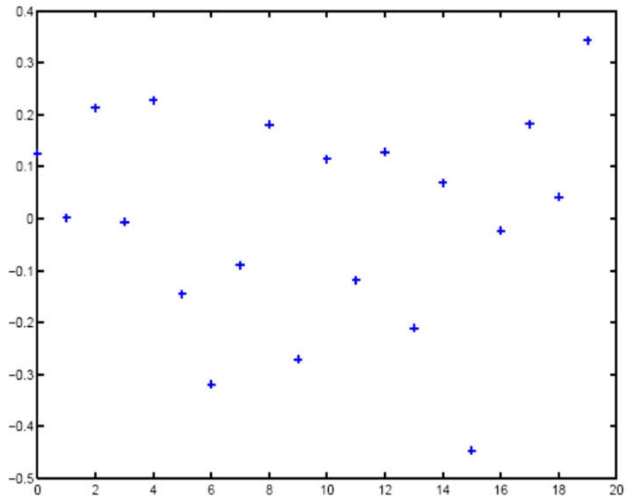


Screening Residuals Helps Determine Correct Model





Model 1:
$Y_{t_i} = ae^{\alpha t_i} + \epsilon_i, \epsilon_i \sim iid\ N(0, \sigma^2)$



Model 2:
$\log Y_{t_i} = \log a + \alpha t_i + \epsilon_i, \epsilon_i \sim iid\ N(0, \sigma^2)$

Model 1 is not compatible with the model assumption (noise terms are iid normal).

3. Palm Calculus, PASTA, Importance of the Sampling Method

Who says the truth ?

SovRail: according to our systematic tracking system, probability of a train being late $\leq 5\%$

BorduKonsum: according to our consumer survey, probability of being late $\approx 30\%$

Palm Calculus

Palm's inversion formula compares distributions seen with different sampling methods (example: seen by an arriving customer vs at an arbitrary point in time). [Brémaud] [Le Boudec].

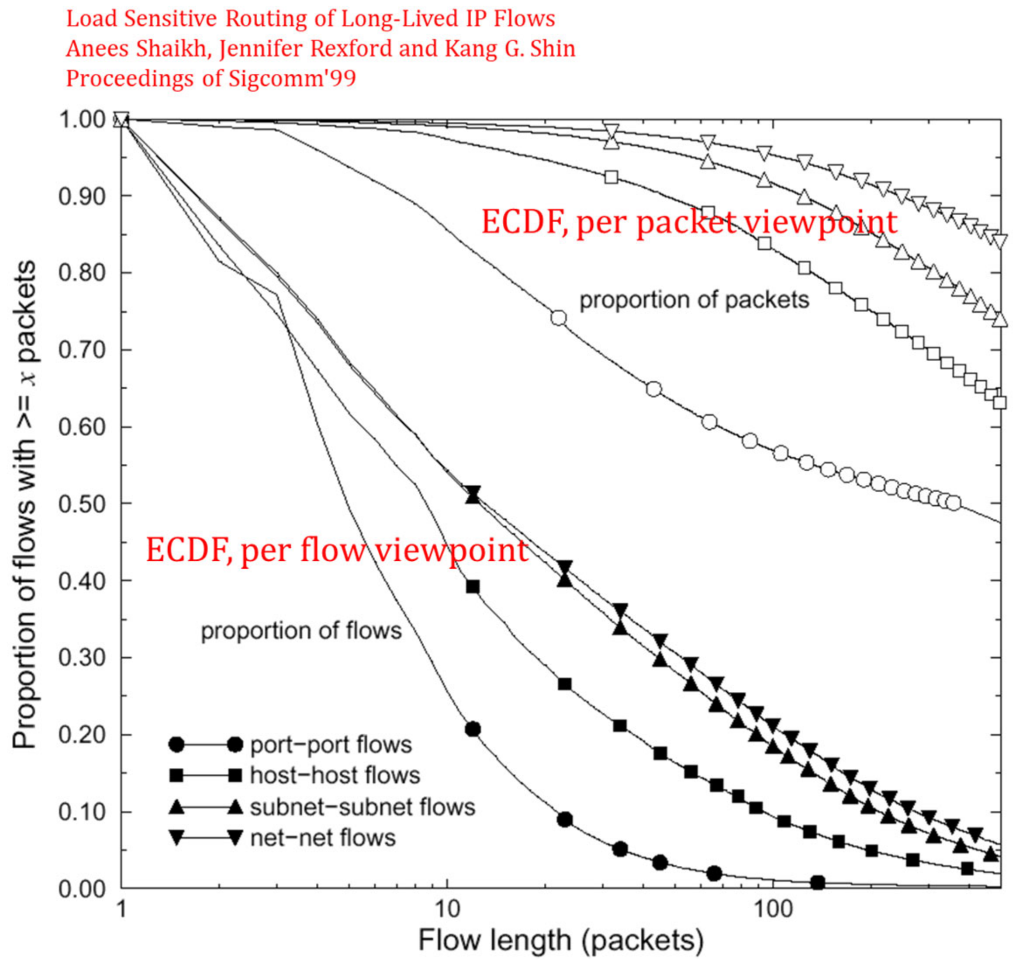
Applies well beyond queuing theory.

Example:

f_F: PDF of flow size sampled per flow

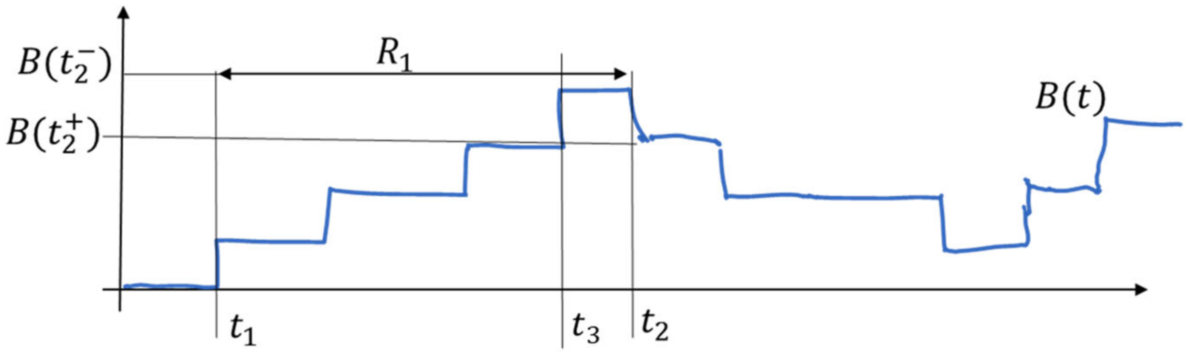
f_P: PDF of flow size sampled per packet

$$f_P(s) = \eta s \, f_F(s)$$



Simulation View of Palm Calculus

Imagine a simulation and compare the statistics obtained with different viewpoints.



Example: Little's formula $N = \lambda R$

Simulator updates counters responseTimeCtr $= \sum_{n=1}^{\text{nbCust}} R_n$ and backlogCtr $= \int_0^T B(t)dt$

At every event, do responseTimeCtr $+= \ (t_2 - t_3)B(t_2^-)$ and backlogCtr $+= \ (t_2 - t_3)B(t_2^-)$, i.e same updates, i.e both counters are equal.

At simulation end do

$R =$ responseTimeCtr$/\text{nbCust}$, $\lambda = \text{nbCust}/T$ and $N =$ backlogCtr$/T$

Thus $\lambda\, R = \frac{\text{nbCust}}{T} \times \frac{\text{responseTimeCtr}}{\text{nbCust}} = \frac{\text{responseTimeCtr}}{T} = N$

Who Says the Truth?

Imagine a simulation with N arrival events.

$D_n = 1_{\text{event } n \text{ is late}}$

$P_n =$ number of passengers leaving train at event n

Sovrail estimate's is $\bar{D} = \frac{1}{N}\sum_{n=1}^{N} D_n$

BorduKonsum's estimate is $D^* = \frac{\sum_{n=1}^{N} P_n D_n}{\sum_{n=1}^{N} P_n}$

Thus $D^* = \bar{D}\,\frac{\bar{P}_{\text{late}}}{\bar{P}}$ with $\bar{P} = \frac{1}{N}\sum_{n=1}^{N} P_n$ and $\bar{P}_{\text{late}} = \frac{\sum_{n=1}^{N} P_n D_n}{\sum_{n=1}^{N} D_n}$

If there are 6 times more passengers in late trains, both estimations are compatible !

SovRail: according to our systematic tracking system, probability of a train being late $\lesssim 5\%$

BorduKonsum: according to our consumer survey, probability of being late $\approx 30\%$

Conclusion

Simulators as thought experiments help students

- remove the magic from probabilistic statements;
- understand and apply classical estimation theory;
- discover Palm calculus formulas.

References

[Barabnasi] Barabási, A.L., 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1987), p.20120375.

[Brémaud] Brémaud, P., 2020. *Point Process Calculus in Time and Space: An Introduction with Applications* (Vol. 98). Springer Nature.

[Le Boudec] Le Boudec, J.Y. *Performance evaluation of computer and communication systems*. EPFL Press 2011, https://leboudec.github.io/perfeval

[Pishro-Nik] H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at https://www.probabilitycourse.com, Kappa Research LLC, 2014.

[Weber] R. Weber, course on Probability for first year mathematicians at Cambridge in winter 2015, http://www.statslab.cam.ac.uk/~rrw1/prob/prob-weber.pdf