

Performance Modeling of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

Sample statistics

Università degli studi di Roma Tor Vergata
Department of Civil Engineering and Computer Science Engineering

Copyright © Vittoria de Nitto Personè, 2021
<https://creativecommons.org/licenses/by-nc-nd/4.0/>



1

Discrete-Event simulation
Sample statistics

- Simulation involves *a lot* of data
- Must “compress” the data into meaningful statistics
- Collected data is a *sample* from a much larger *population*
- Two types of statistical analysis
 - “Within-the-run”
 - “Between-the-runs” (replication)
- Essence of statistics: analyze a sample and draw inferences

Il campione deve ben rappresentare una generalità

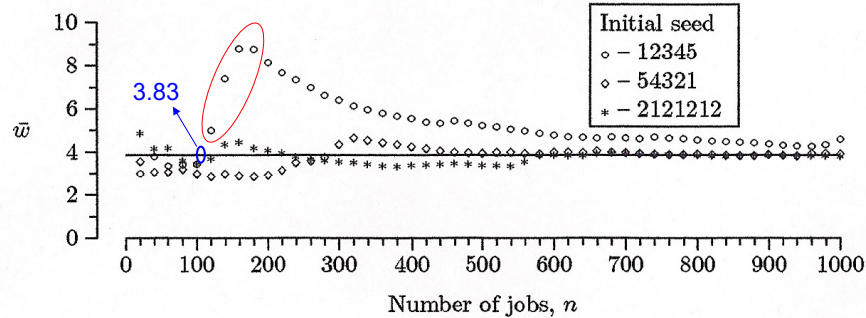
Prof. Vittoria de Nitto Personè

2

2

- The accumulated average wait was printed every 20 job

stiamo vedendo i tempi di risposta, * è tempo risposta dei job su asse x.



- The convergence of w is slow, erratic, and dependent on the initial seed

campionamento fatto ogni 20, se faccio 3 run devo farlo con 3 semi diversi.

Sto "guardando" 3 scenari diversi per quella coda, che hanno arrivi, servizi dello stesso tipo. Cambiano i valori in sè.

Prof. Vittoria de Nitto Personè

3

abbiamo * e °,
che sono run diverse,
a causa del seed preso.

dopo si converge al valore
teorico.

3

arrival and service processes are **uncoupled**

stream 0 for arrivals, stream 1 for services

for 10025 jobs

average interarrival time = 1.99
average wait = 3.92
average delay = 2.41
average service time = 1.50
average # in the node ... = 1.96
average # in the queue .. = 1.21
utilization = 0.75

sugli stessi arrivi voglio
esaminare servizi diversi?
cambio stream per servizi,
passo da 1 a 2.
come vediamo abbiamo dei leggeri
cambiamenti.

stream 0 for arrivals, stream 2 for services (or e.g. stream 128 to get more separation)

for 10025 jobs

average interarrival time = 1.99
average wait = 3.86
average delay = 2.36
average service time = 1.50
average # in the node ... = 1.93
average # in the queue .. = 1.18
utilization = 0.75

Theoretical values

\bar{r}	\bar{w}	\bar{d}	\bar{s}	\bar{l}	\bar{q}	\bar{x}
2.00	3.83	2.33	1.50	1.92	1.17	0.75

Prof. Vittoria de Nitto Personè

4

4

Sample Mean and Standard Deviation

Consider a sample x_1, x_2, \dots, x_n (continuous or discrete), let us define:

- *sample mean* $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- *sample variance* $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
- *sample standard deviation* $s = \sqrt{s^2}$
- *coefficient of variation* s / \bar{x}

Prof. Vittoria de Nitto Personè

5

5

- *mean*: a measure of central tendency misura tendenza centrale.

- *variance and deviation*: measures of dispersion about the mean
variabilità rispetto a quella tendenza.

easier math
(no square root)

same units as data,
mean

- note that *coefficient of variation (CV)* is unit-less, but a common shift in data changes the CV

Prof. Vittoria de Nitto Personè

6

6

Relating the mean and standard deviation

Consider the root-mean-square (rms) function

$$d(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - x)^2}$$

questa funzione è la dispersione rispetto ad un qualsiasi valore 'x'.
se ci metto $x = E[x]$, cioè la media, tendenza centrale,
questo valore minimizza la funzione $d(x)$.

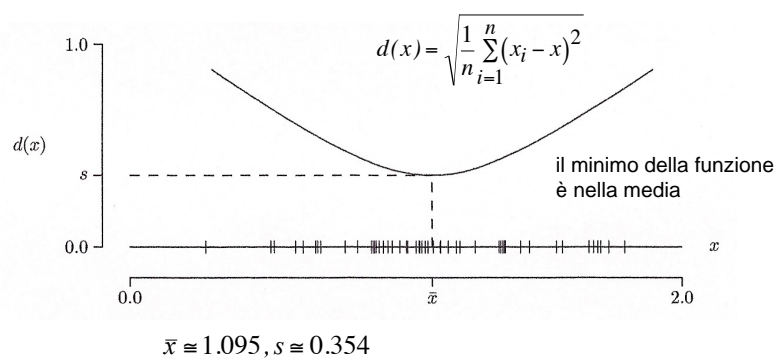
- $d(x)$ measures dispersion about any value x
- the mean \bar{x} gives the smallest possible value for $d(x)$ (Theorem 4.1.1)
- The standard deviation s is that smallest value

Prof. Vittoria de Nitto Personè

7

7

50 samples from program buffon



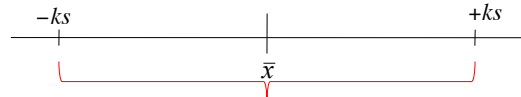
Prof. Vittoria de Nitto Personè

8

8

Chebyshev's inequality

Consider the number of points that lie within k standard deviations of the mean



- Points farthest from the mean make the most contribution to s

Define the set $S_k = \{x_i | \bar{x} - ks < x_i < \bar{x} + ks\}$

Let $p_k = |S_k|/n$ be the proportion of x_i within $\pm ks$ of \bar{x}
empirica

$$p_k \geq 1 - \frac{1}{k^2} \quad (k > 1)$$

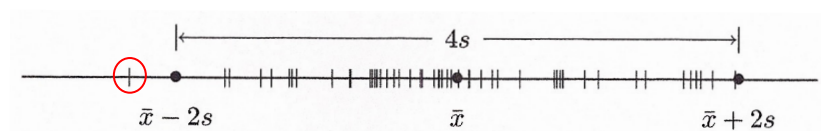
Prof. Vittoria de Nitto Personè

9

9

Chebyshev's inequality

- for any sample, at least 75% of the points lie within $\pm 2s$ of \bar{x}
- for $k=2$, the inequality is very conservative:
typically 95% lie within $\pm 2s$ of \bar{x}
- $\bar{x} \pm 2s$ defines the "effective width" of a sample



- most (but not all) points will lie in this interval
- **outliers** should be viewed with suspicion

Prof. Vittoria de Nitto Personè

10

10

Linear data transformations

- Often need to convert to different units after data has been collected
- let x'_i be the "new data": $x'_i = ax_i + b$

qui sono tutte CAMPIONARIE

- **sample mean**

la media viene trasformata in modo lineare.

$$\bar{x}' = \frac{1}{n} \sum_{i=1}^n x'_i = \frac{1}{n} \sum_{i=1}^n (ax_i + b) = \frac{a}{n} \left(\sum_{i=1}^n x_i \right) + b = a\bar{x} + b$$

- **sample variance** $(s')^2 = \frac{1}{n} \sum_{i=1}^n (x'_i - \bar{x}')^2 = \dots = a^2 s^2$

- **sample standard deviation** $s' = |a|s$

Prof. Vittoria de Nitto Personè

11

11

Examples of Linear Data Transformations

- suppose x_1, x_2, \dots, x_n measured in seconds
 - to convert to minutes, let $x'_i = x_i/60$

($a=1/60, b=0$)

se media = 45, la trasformazione è:

$$\bar{x}' = \frac{45}{60} = 0.75 \qquad s' = \frac{15}{60} = 0.25 \quad (\text{minutes})$$

- **standardize data**
($a=1/s, b=-\bar{x}/s$)

$$x'_i = \frac{1}{s} x_i - \frac{\bar{x}}{s}$$

$$x'_i = \frac{x_i - \bar{x}}{s}$$

Then

$$\bar{x}' = 0$$

$$s' = 1$$

Used to avoid problems with very large (or small) valued samples

Prof. Vittoria de Nitto Personè

12

12

per trattare dati molto grandi o molto piccoli. I coefficienti hanno al denominatore la deviazione standard. Ciò ci permette di avere media campionaria standardizzata = 0 e std deviation = 1. Allora nella disuguaglianza di Chebyshev, nell'intorno di 2 c'è l'intero campione.

Nonlinear data transformations

Quando la uso? la uso se, invece di un valore specifico del campione, mi interessa ad una PROPRIETA' di quel campione.

- usually involves a Boolean (two-state) outcome
- the *value* of x_i is not as important as the *effect*
- let A be a fixed set; then

A è un insieme

$$x'_i = \begin{cases} 1 & x_i \in A \\ 0 & \text{otherwise} \end{cases}$$

- let p be the proportion of x_i that fall in A :

$$\text{(media empirica)} \quad p = \frac{\text{the number of } x_i \text{ in } A}{n}$$

then

$$\begin{array}{ll} \bar{x}' = p & s' = \sqrt{p(1-p)} \\ \text{media} & \text{std deviation} \end{array}$$

Non voglio sapere x' quanto vale, piuttosto se cade in un certo range A , allora la trasformazione è 1 se cade in A , 0 altrimenti.

Prof. Vittoria de Nitto Personè

13

13

Examples of Nonlinear Data Transformations

voglio sapere sia tempo attesa medio, sia QUANTI subiscono attesa.

For the single server service queue

- let $x_i = d_i$ be the delay for job i

- let $A = \mathbb{R}^+$, then $x'_i = 1$ iff $d_i > 0$
attesa può essere
qualsiasi tempo positivo

- from exerc. 1.2.3 proportion of job delayed is $p = 0.723$

- then $\bar{x}' = 0.723$ and $s = \sqrt{(0.723)(0.277)} = 0.448$

Prof. Vittoria de Nitto Personè

14

14

Computational considerations

Consider the sample standard deviation equation

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Requires two passes through the data:

1. Compute the mean \bar{x}
2. Compute the squared differences about \bar{x} distanza di ogni campione dalla media

Must store or re-create the entire sample!
bad when n is large!

memorizzo 1 mln di dati? NO
dovrei rigenerarli due volte? una per calcolare la media, e una per la deviazione std?
Posso farlo in uno step?

Prof. Vittoria de Nitto Personè

15

15

The conventional one-pass Algorithm

non dovrebbe essere varianza?
Consider the sample standard deviation equation

$$\begin{aligned} s^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \end{aligned}$$

by separating and simplifying

$$= \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$$

round-off error, overflow perchè ho somma con quadrati dentro.

Prof. Vittoria de Nitto Personè

16

16

Welford's one-pass algorithm

- running sample mean until i

$$\bar{x}_i = \frac{1}{i}(x_1 + x_2 + \dots + x_i)$$

- running sample sum of squared deviations until i

$$v_i = (x_1 - \bar{x}_i)^2 + (x_2 - \bar{x}_i)^2 + \dots + (x_i - \bar{x}_i)^2$$

- \bar{x}_i and v_i can be computed recursively ($\bar{x}_0 = 0, v_0 = 0$) :

$$\bar{x}_i = \bar{x}_{i-1} + \frac{1}{i}(x_i - \bar{x}_{i-1})$$

le calcolo in funzione
del passo precedente.

$$v_i = v_{i-1} + \left(\frac{i-1}{i}\right)(x_i - \bar{x}_{i-1})^2$$

- \bar{x}_n is the sample mean, v_n / n is the variance

Prof. Vittoria de Nitto Personè

17

17

Welford's Algorithm (program uvs)

- No *a priori* knowledge of the sample size n required
- Less prone to accumulated round-off error

$\bar{x}_i = \bar{x}_{i-1} + \frac{1}{i}(x_i - \bar{x}_{i-1})$

```

n = 0;
x_bar = 0.0;
v = 0.0;
while (more data) {
    x = GetData();
    n++;
    d = x - x_bar;
    v = v + d * d * (n - 1) / n;
    x_bar = x_bar + d / n;
}
s = sqrt(v / n);
return n, x_bar, s;

```

Prof. Vittoria de Nitto Personè

18

18

Welford's Algorithm (program `uvs`)

- No *a priori* knowledge of the sample size n required
- Less prone to accumulated round-off error

```

n = 0;
x̄ = 0.0;
v = 0.0;
while (more data) {
    x = GetData();
    n++;
    d = x - x̄;
    v = v + d * d * (n - 1) / n;
    x̄ = x̄ + d / n;
}
s = sqrt(v / n);
return n, x̄, s;

```

$$v_i = v_{i-1} + \left(\frac{i-1}{i}\right)(x_i - \bar{x}_{i-1})^2$$

Prof. Vittoria de Nitto Personè

19

19

Example

- let x_1, x_2, \dots, x_n be $Uniform(a, b)$ random variates

- in the limit as $n \rightarrow \infty$

$$\bar{x} \rightarrow \frac{a+b}{2} \qquad s \rightarrow \frac{b-a}{\sqrt{12}}$$

media teorica, e deviazione standard teorica

- using $Uniform(0, 1)$ \bar{x} and s should converge to

$$\frac{0+1}{2} = 0.5 \qquad \frac{1-0}{\sqrt{12}} \approx 0.2887$$

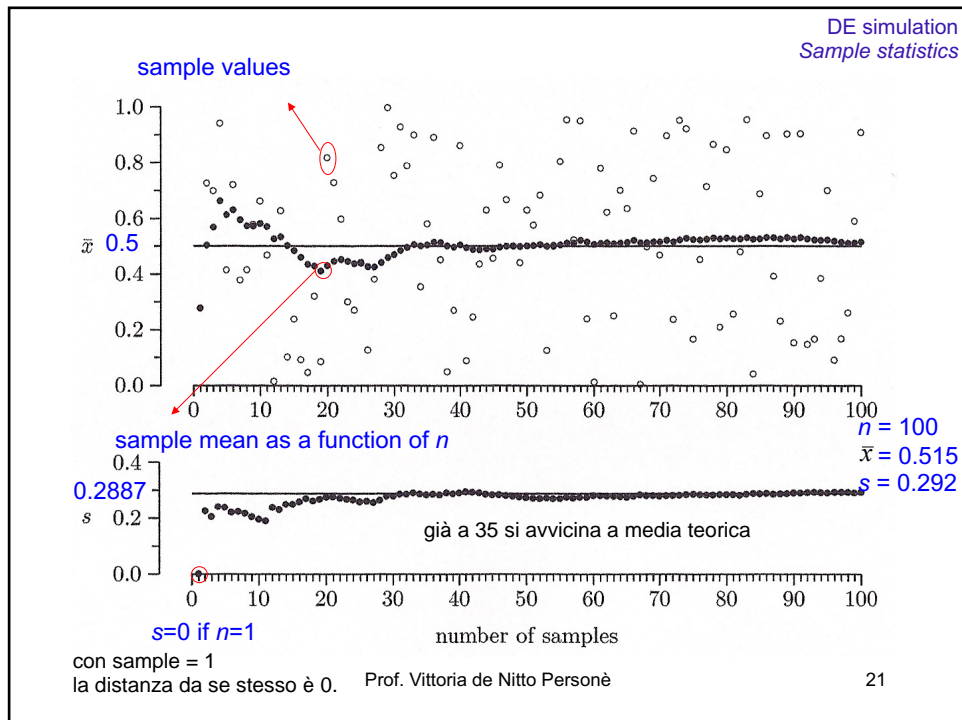
- The convergence to theoretical values is not necessarily monotone

il campione, se piccolo, non arriva subito a questi valori teorici, ma se n cresce dovrebbe arrivarci.

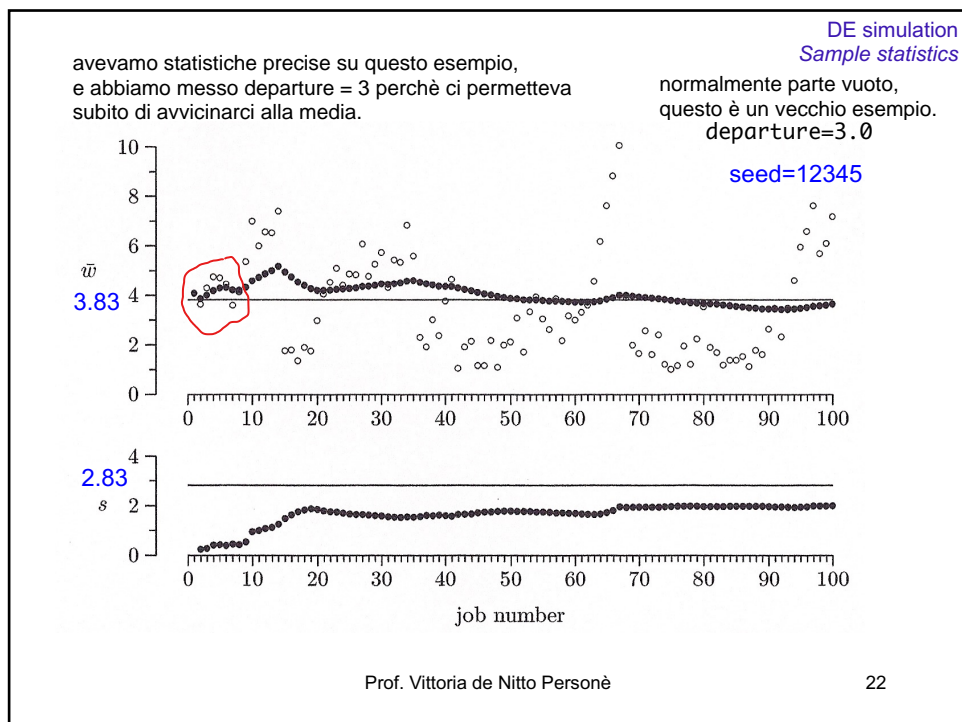
Prof. Vittoria de Nitto Personè

20

20



21



22

Nel primo grafico dei tempi di risposta, c'è correlazione perchè a gruppi i job attendono un tempo sopra la media (all'inizio, quelli più vicini in rosso sono simulati). La coda si forma, i tempi in coda si allungano sopra 3.83, poi sistema si svuota, o tempi richiesti sono inferiori (sotto 3.83), poi si rialzano etc...

Il campione della slide sopra non è correlato, anche ad occhio non osserviamo pattern.

La deviazione standard di del caso in QUESTA SLIDE è sottostimato, e quindi dovremmo AGGIUSTARE QUESTO BIAS.

time-averaged sample statistics

Il campione sample-path (cioè funziona coi sample) che osserviamo (processo stocastico in funzione di t) è soggetto a:

- Let $x(t)$ be the sample path of a stochastic process for $0 < t < \tau$

- Sample-path mean $\bar{x} = \frac{1}{\tau} \int_0^\tau x(t) dt$

- Sample-path variance $s^2 = \frac{1}{\tau} \int_0^\tau (x(t) - \bar{x})^2 dt$

- Sample-path standard deviation $s = \sqrt{s^2}$

- One-pass equation for variance
$$s^2 = \left(\frac{1}{\tau} \int_0^\tau x^2(t) dt \right) - \bar{x}^2$$

$$s^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$$

Prof. Vittoria de Nitto Personè

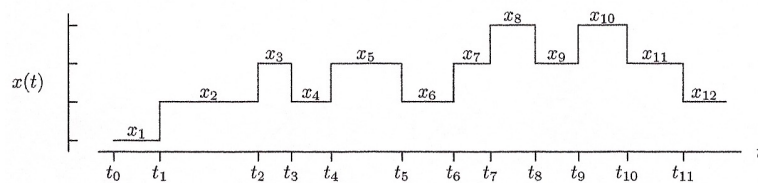
23

23

Anche loro sono soggetti ad essere stepwise.

Computational considerations

- For DES, a sample path is *piecewise constant*
- Changes in the sample path occur at *event times* t_0, t_1, \dots



- For computing statistics, integrals reduce to summations

Prof. Vittoria de Nitto Personè

24

24

Computational sample-path formulas

Consider a piecewise constant sample path

$$x(t) = \begin{cases} x_1 & t_0 < t \leq t_1 \\ x_2 & t_1 < t \leq t_2 \\ \vdots & \vdots \\ x_n & t_{n-1} < t \leq t_n \end{cases}$$

Da stepwise,
nel continuo passiamo
agli integrali.

- **Sample-path mean** $\bar{x} = \frac{1}{\tau} \int_0^\tau x(t) dt = \frac{1}{t_n} \sum_{i=1}^n x_i \delta_i$ with $\delta_i = t_i - t_{i-1}$
inter-event time
- **Sample-path variance**

$$s^2 = \frac{1}{\tau} \int_0^\tau (x(t) - \bar{x})^2 dt = \frac{1}{t_n} \sum_{i=1}^n (x_i - \bar{x})^2 \delta_i = \left(\frac{1}{t_n} \sum_{i=1}^n x_i^2 \delta_i \right) - \bar{x}^2$$

Prof. Vittoria de Nitto Personè

25

25

Welford's sample path Algorithm

- based on the definitions

$$\bar{x}_i = \frac{1}{i} (x_1 \delta_1 + x_2 \delta_2 + \dots + x_i \delta_i)$$

$$v_i = (x_1 - \bar{x}_i)^2 \delta_1 + (x_2 - \bar{x}_i)^2 \delta_2 + \dots + (x_i - \bar{x}_i)^2 \delta_i$$

- \bar{x}_i is the sample-path mean of $x(t)$ for $t_0 \leq t \leq t_i$
- v_i / t_i is the sample-path variance
- \bar{x}_i and v_i can be computed recursively ($\bar{x}_0 = 0, v_0 = 0$)

$$\bar{x}_i = \bar{x}_{i-1} + \frac{\delta_i}{t_i} (x_i - \bar{x}_{i-1})$$

$$v_i = v_{i-1} + \frac{\delta_i t_{i-1}}{t_i} (x_i - \bar{x}_{i-1})^2$$

Prof. Vittoria de Nitto Personè

26

26

Exercises

- Exercises: 4.1.7, 4.1.8