Computer system modeling and simulation

6. Simulation output analysis

Sosina M.
Addis Ababa institute of technology (AAiT)
2012 E.C.

Input modeling

- In real-world simulation applications, estimating the appropriate distribution for input parameters is a major task
- Development of a useful model of input data
  - Collect data from the real system of interest
  - Identify a probability distribution to represent the input process
  - Choose parameters that determine a specific instance of the distribution family
    - The parameters can be estimated from the data, if available
  - Evaluate the chosen distribution and the associated parameters for goodness of fit
    - chi-square and the Kolmogorov-Smirnov tests are standard goodness-of-fit tests

Output analysis

- Output analysis is the examination of data generated by a simulation
  - to predict the performance of a system
  - to compare the performance of two or more alternative system designs
- The Simulation
  - Receives as inputs parameters represented by random variables
  - produces outputs
- Compute mean value and an estimation error

Type of simulation

- **Terminating simulation**
  - Runs for some duration of time T_E , where E is a specified event or set of events that stops the simulation
    - Starts with a well specified initial conditions
    - Ends at stopping time T_E
  - Dependency from the initial conditions
- **Non-terminating simulation**
  - Runs continuously, or at least over a very long period of time
  - Initial conditions and simulation period defined by the analyst
  - Study the steady state properties of the system – properties that are not influenced by the initial condition of the model

Stochastic nature of output data

- Model outputs consist of one or more random variables

- ***Example- M/G/1 queueing***
  - Poisson arrival rate = 0.1 per minute and service time $\mu = 9.5$, $s = 1.75$
  - System performance: long-run mean queue length, $L_Q(t)$
  - Suppose we run a single simulation for a total of 5000 minutes
    - Divide the time interval [0, 5000) into 5 equal subintervals of 1000 minutes
    - Run 3 independent replications

Stochastic nature of output data

❑ *Example- M/G/1 queueing*

Batching Interval (Minutes)	*Batch, j*	*Replication* 1, Y_{1j}	2, Y_{2j}	3, Y_{3j}
[0, 1000)	1	3.61	2.91	7.67
[1000, 2000)	2	3.21	9.00	19.53
[2000, 3000)	3	2.18	16.15	20.36
[3000, 4000)	4	6.92	24.53	8.11
[4000, 5000)	5	2.82	25.19	12.62
[0, 5000)		$\bar{Y}_{1\cdot} = 3.75$	$\bar{Y}_{2\cdot} = 15.56$	$\bar{Y}_{3\cdot} = 13.66$

- variability in stochastic simulation both within a single replication and across different replications
- The average across 3 replications can be regarded as independent observations, but averages within a replication are not

Measures of performance and their estimation

- **Point estimation**
- The point estimator of θ based on the data $\{Y_1, \dots, Y_n\}$
- Discrete time data - $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ where $\hat{\theta}$ =sample mean on a sample size n
- Continuous time data - $\hat{\theta} = \frac{1}{T_E}\int_0^{T_E} Y(t)dt$
- The point estimator $\hat{\theta}$ is said to be unbiased for θ if

$$E(\hat{\theta}) = \theta$$

Confidence interval estimation

- An estimation of the error we make when we estimate the average value X of a random quantity using an averaging process over a limited number of samples and replications

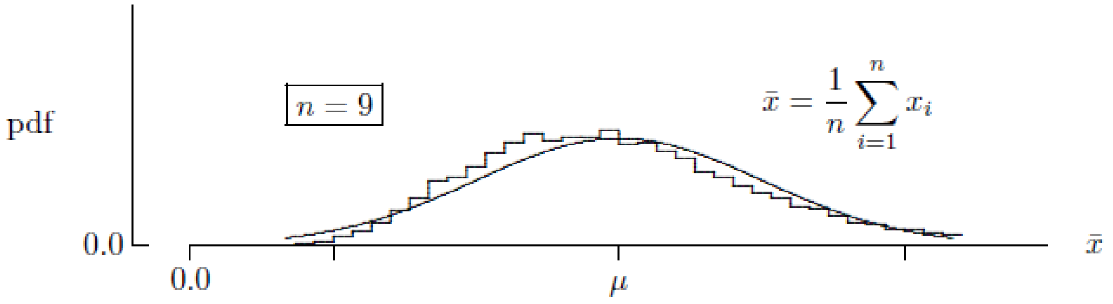
- *Hypothesis*
  - The observation process is stationary
  - X has a mean μ and standard deviation σ
  - n independent observations of X (x1, x2,…,xn)

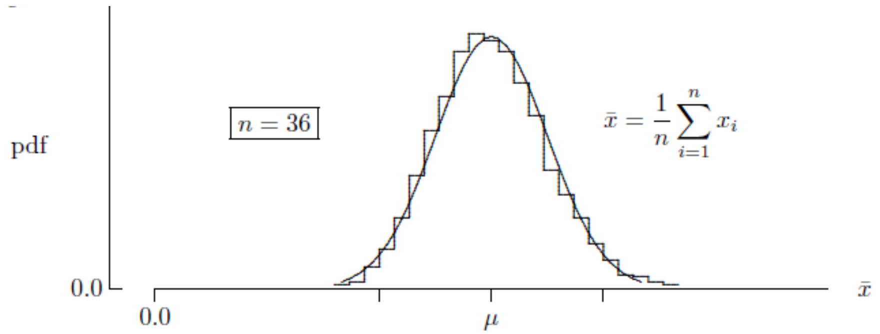
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - Then the mean of E($\bar{X}$)= μ and SD($\bar{X}$)= ${}^{\sigma}/_{\sqrt{n}}$
  - Central limit theorem – for large n, the estimate is normally distributed

Confidence interval estimation

- Example- 10000 Exponential random variate samples of size n = 9 and 10 000 samples of size n = 36



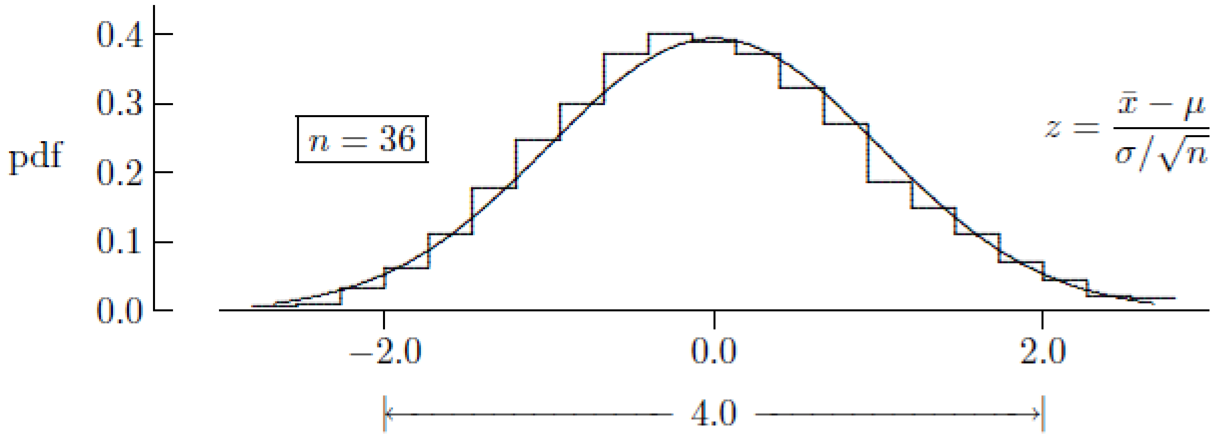


Confidence interval

- Transform N(μ, $^{\sigma}/_{\sqrt{n}}$) to N(0,1)

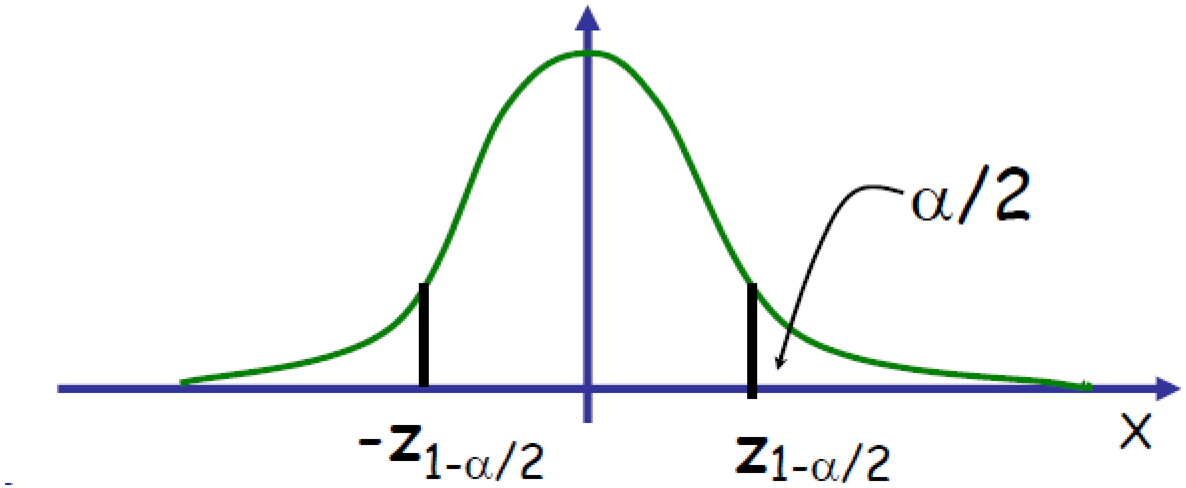
$$z = \frac{x-\mu}{^{\sigma}/_{\sqrt{n}}}$$

  - Z is used to evaluate the error in the estimation
- Example



Confidence intervals

- The probability that z is in the interval $[-z_{1-\alpha/2}, z_{1+\alpha/2}]$ is (1- α)



- $P\left\{\bar{x} - z_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha$
- $1-\alpha$=the confidence level

Confidence interval

- The central limit theorem is valid only for large n (>30)

- For n<30, student t distribution with n-1 degrees of freedom is used

$$P\left\{\bar{x} - t_{\alpha/2,\, n-1}\,\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,\, n-1}\,\frac{s}{\sqrt{n}}\right\} = 1 - \alpha$$

- $t_{\alpha/2,\, n-1}\,\frac{s}{\sqrt{n}}$ – is a measure of error

Example

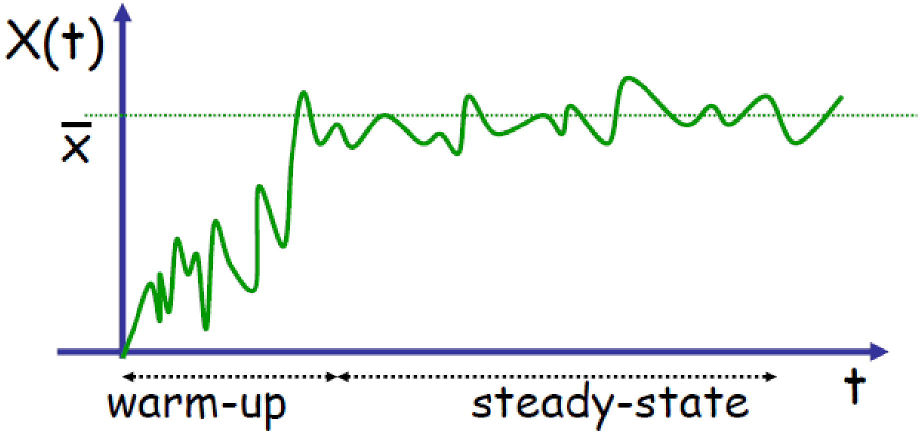
- 9 simulation runs provided the following results

  $$\bar{x} = 65,\ E(\sum_{i=1}^{9}(x-\bar{x})^2) = 3560$$

  - Compute 90% and 99% confidence intervals
    - Variance =3560/8=445
    - 90% confidence interval
      - ✓ Student t with 8 degrees of freedom
      - ✓ 1- α=0.9, $\frac{\alpha}{2} = 0.05$
      - ✓ Use the value $t_{0.05,\,8}$
      - ✓ $I_{0.9} =$ [51.92,78.08]

Independent observation

- The theory of confidence intervals assumes
  - The process is stationary
  - The observations are independent
- Identifying the stationarity conditions – removing the transient
- Long simulation runs - to collect significant samples
- Several independent simulation runs – to collect independent samples



Removing warm-up transient

- Visual inspection and heuristics to identify to warm-up period
  - the variance of the measures during the transient is higher than at steady-state
- Solutions
  - Long runs: running a simulation for a long time, the effect of the warp-up phase over the performance measures is reduced
  - Selection of initial conditions near to the steady-state conditions
  - Initial data removal
    - Given n observations, remove the first k observations
    - removal of samples collected during the warm-up transient will change the average of the remaining data, while removing samples during the steady-state will not influence too much the average

Initial data removal

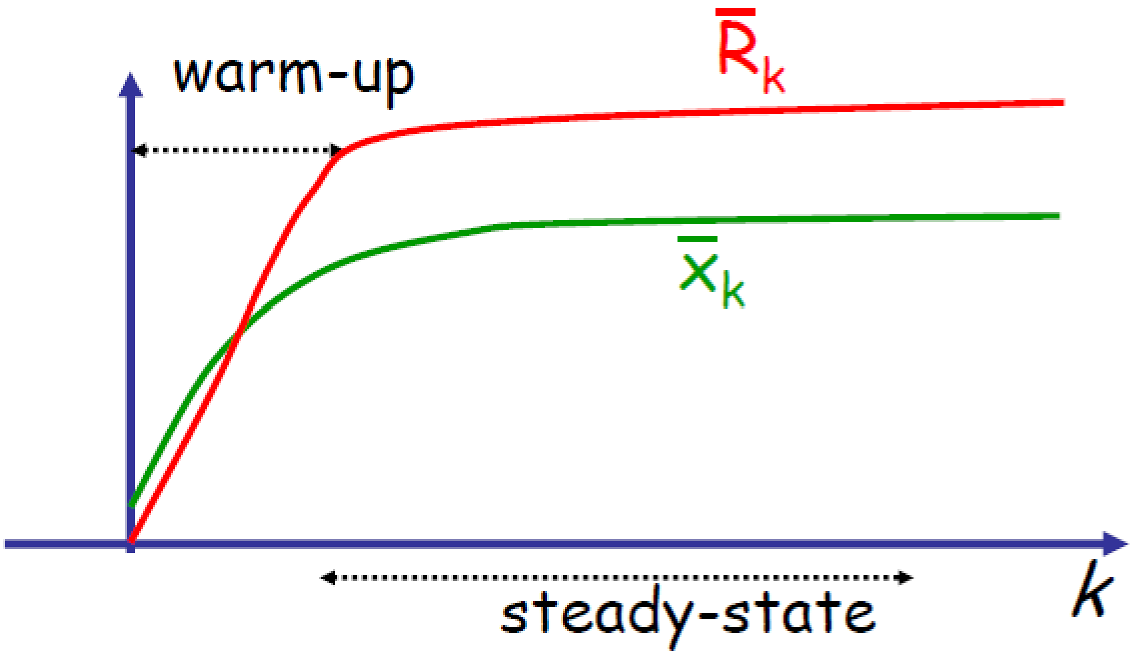
- Compute the mean

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\overline{x_k} = \frac{1}{n-k}\sum_{j=k+1}^{n} x_j$$

- Compute the relative variation

$$R_k = \frac{\overline{x_k} - \bar{x}}{\bar{x}}$$



- After drawing a graph of R_k as a function of k, choose the value of k