MEM6810 Engineering Systems Modeling and Simulation
工程系统建模与仿真

Theory Analysis

Lecture 9: Output Analysis II: Comparison

SHEN Haihui 沈海辉

Sino-US Global Logistics Institute
Shanghai Jiao Tong University

shenhaihui.github.io/teaching/mem6810f
shenhaihui@sjtu.edu.cn

Spring 2022 (full-time)

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY
董浩云航运与物流研究院 CY TUNG Institute of Maritime and Logistics
中美物流研究院（工程系统管理研究院） Sino-US Global Logistics Institute (Institute of Industrial & System Engineering)

Contents

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

1 Introduction

2 Comparison of Two Designs
- Significant Difference
- Independent Sampling
- Common Random Numbers

3 Comparison of Multiple Designs
- Bechhofer's Procedure
- Paulson's Procedure
- Ranking and Selection Review
- Multi-Arm Bandit Problem

Introduction

- We have learnt how to estimate the *absolute performance* of a simulation model.
- We now discuss how to compare two or more simulation models, i.e. to estimate their *relative performance*.
- Here, different simulation models may refer to different designs, operation policies, etc., of a simulated system; in this lecture we simply call them *different (system) designs*.
- It is one of the most important uses of simulation.

Introduction

- **Key Question:** Are the observed differences due to
  - the actual differences on the expected performance of system designs?
  - or the random errors in the simulation outputs?
- The comparison can be classified into two types:
  - Two system designs: using confidence interval of the difference.
  - Multiple (more than two) system designs: selection of the best.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

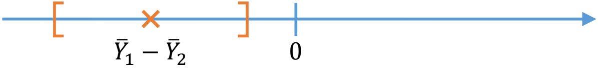
- Let θ_1 and θ_2 be the mean performance of the two system designs in simulation.
- To compare θ_1 and θ_2, we simply construct the point and interval estimates of $\theta_1 - \theta_2$
- Suppose we have the simulation output data from simulation of two system designs.†

System	*Replication* 1	2	$\cdots$	R_i	*Sample Mean*	*Sample Variance*
1	Y_{11}	Y_{21}	$\cdots$	$Y_{R_1 1}$	$\bar{Y}_1$	S_1^2
2	Y_{12}	Y_{22}	$\cdots$	$Y_{R_2 2}$	$\bar{Y}_2$	S_2^2

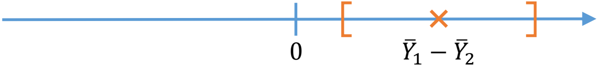
- Point estimator of $\theta_1 - \theta_2$: $\bar{Y}_1 - \bar{Y}_2$.
- Approximate $1-\alpha$ CI: $\bar{Y}_1 - \bar{Y}_2 \pm t_{v,1-\alpha/2} \times \mathrm{s.e.}(\bar{Y}_1 - \bar{Y}_2)$.
    - $\mathrm{s.e.}(\bar{Y}_1 - \bar{Y}_2)$ is the estimator of standard error of $\bar{Y}_1 - \bar{Y}_2$; see more details about this quantity and v later.

† The notation here is different from that in Lec 7; the second subscript indicates different system designs.

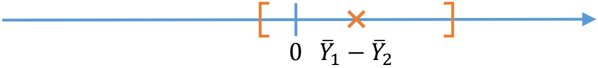
- Case 1 – Strong evidence that $\theta_1 < \theta_2$:



- Case 2 – Strong evidence that $\theta_1 > \theta_2$:



- Case 3 – No strong evidence that one is larger than the other:



  - It does not imply $\theta_1 = \theta_2$!

Comparison of Two Designs

- The first two cases are conclusive.
- If in case 3, then we increase the number of replications R_1 and/or R_2, after which the CI would likely shift, and definitely shrink in length.
- We will shrink the CI until case 1 or 2 is achieved, or the confidence interval is so narrow, which suggests that we do not need to separate them.

- For the comparison of performance of two designs, there is an important distinction between
  - *statistically significant difference* (统计意义上的显著区别);
  - *practically significant difference* (实际意义上的显著区别).
- Statistical significance answers the following questions:
  - Is the observed difference $\bar{Y}_1 - \bar{Y}_2$ larger than its variability?
  - Have we collected enough data to be confident that the observed difference is real (not just by chance)?
- Practical significance answers the following question:
  - Is the true difference $|\theta_1 - \theta_2|$ large enough so it is worthwhile to separate them?

- Cases 1 and 2 imply a statistically significant difference, while case 3 does not.
- In case 1, we may reach the conclusion that $\theta_1 < \theta_2$ and decide that design 2 is better (suppose larger is better).
- However, if the actual difference $|\theta_1 - \theta_2|$ is very small, then it might not be worth the cost to replace design 1 with design 2.
- Confidence intervals do not answer the question of practical significance directly.
  - Instead, they bound, with probability $1 - \alpha$, the true difference $\theta_1 - \theta_2$ within the range $\bar{Y}_1 - \bar{Y}_2 \pm t_{v,\,1-\alpha/2} \times \mathrm{s.e.}(\bar{Y}_1 - \bar{Y}_2)$.
  - Whether a difference within these bounds is practically significant depends on the particular problem.

- Independent sampling means that different random number streams are used to simulate the two systems.
  - All the observations of system 1 $\{Y_{r1} : r = 1, \dots, R_1\}$ are statistically independent of all the observations of system 2 $\{Y_{r2} : r = 1, \dots, R_2\}$.
- Suppose $\mathrm{Var}(Y_{r1}) = \sigma_1^2$ and $\mathrm{Var}(Y_{r2}) = \sigma_2^2$. Due to the independence,

$$\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2) = \mathrm{Var}(\bar{Y}_1) + \mathrm{Var}(\bar{Y}_2) = \frac{\sigma_1^2}{R_1} + \frac{\sigma_2^2}{R_2}.$$

  - Standard error of $\bar{Y}_1 - \bar{Y}_2$ is $\sqrt{\frac{\sigma_1^2}{R_1} + \frac{\sigma_2^2}{R_2}}$.
- σ_i^2 is estimated via sample variance

$$S_i^2 = \frac{1}{R_i - 1} \sum_{r=1}^{R_i} (Y_{ri} - \bar{Y}_i)^2.$$

  - Standard error of $\bar{Y}_1 - \bar{Y}_2$ is estimated via

$$\mathrm{s.e.}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{S_1^2}{R_1} + \frac{S_2^2}{R_2}}. \quad (1)$$

- The $1-\alpha$ CI is approximated by

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{v,\,1-\alpha/2} \times \mathrm{s.e.}(\bar{Y}_1 - \bar{Y}_2). \tag{2}$$

where $\mathrm{s.e.}(\bar{Y}_1 - \bar{Y}_2)$ is given in (1), and the degree of freedom v is

$$v = \frac{[S_1^2/R_1 + S_2^2/R_2]^2}{[S_1^2/R_1]^2/(R_1-1) + [S_2^2/R_2]^2/(R_2-1)}.$$

- The approximated CI (2) is called the *Welch confidence interval* (Welch 1938).
  - Sometimes, people will round v to integer for convenience.

- If $R_1 = R_2 = R$, or we are willing to discard some observations from the system design on which we actually have more data, we can pair Y_{r1} with Y_{r2} to define $Z_r = Y_{r1} - Y_{r2}$, for $r = 1, \ldots, R$.
- Point estimator of $\theta_1 - \theta_2$: $\bar{Z} = \frac{1}{R}\sum_{r=1}^{R} Z_r = \bar{Y}_1 - \bar{Y}_2$.

$$\begin{aligned}\mathrm{Var}(\bar{Z}) &= \frac{\mathrm{Var}(Z_r)}{R} = \frac{\mathrm{Var}(Y_{r1} - Y_{r2})}{R} = \frac{\sigma_1^2 + \sigma_2^2}{R} \\ &= \mathrm{Var}(\bar{Y}_1 - \bar{Y}_2) = \mathrm{Var}(\bar{Y}_1) + \mathrm{Var}(\bar{Y}_2) = \frac{\sigma_1^2 + \sigma_2^2}{R}.\end{aligned} \tag{3}$$

- To estimate $\mathrm{Var}(Z_r)$, instead of estimating σ_1^2 and σ_2^2 separately, we can directly use

$$S^2 = \frac{1}{R-1}\sum_{r=1}^{R}(Z_r - \bar{Z})^2. \tag{4}$$

- Approximate $1-\alpha$ CI:

$$\bar{Z} \pm t_{R-1,\,1-\alpha/2}\frac{S}{\sqrt{R}}. \tag{5}$$

- Common Random Numbers (CRN, also known as correlated sampling): For each replication, the same random numbers are used to simulate both systems.
  - For each replication r, the two estimates, Y_{r1} and Y_{r2}, are correlated.
  - In this case, R_1 and R_2 must be equal, say, $R_1 = R_2 = R$.
- The purpose of using CRN is to induce a positive correlation between Y_{r1} and Y_{r2} for each r and thus to achieve a variance reduction in the point estimator of $\theta_1 - \theta_2$, $\bar{Z}$.

$$\mathrm{Var}(\bar{Z}) = \frac{\mathrm{Var}(Y_{r1} - Y_{r2})}{R} = \frac{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}{R}. \tag{6}$$

- $\mathrm{Var}(\bar{Z})$ in (6) is smaller than that in (3) $\Longrightarrow$ higher precision of point estimator.
- CI is still computed via (4) and (5), but the width will be smaller $\Longrightarrow$ higher precision.

- It is never enough to simply use the same seed for the random-number generator(s):
  - The random numbers must be synchronized: each random number used in one model for some purpose should be used for the same purpose in the other model.
  - E.g., if the ith random number is used to generate a service time at work station 2 for the 5th arrival in model 1, the ith random number should be used for the very same purpose in model 2.
- The CRN idea is also used when we validate simulation model via input-output transformation, where we prefer to compare the model and actual system under the same historical input, rather than generate the input from input model.

SHANGHAI JIAO TONG UNIVERSITY

Comparison of Multiple Designs

- Suppose there are $k > 2$ system designs in total.
- The interested mean performance of design i is θ_i (unknown).
- Some possible goals:
  1. Estimation of each parameter θ_i.
  2. Comparison of each θ_i to a control, say, θ_1 (θ_1 can represent the mean performance of an existing system).
  3. All pairwise comparisons.
  4. Selection of the best θ_i (largest or smallest).
- The first three can be achieved by **simultaneous** construction of confidence intervals, whereas the last by some **selection approaches**.
- From now on, without loss of generality, let's *assume the best* θ_i *is the largest one*.

- Assumption 1: For each design i with mean performance θ_i, the noisy output $Y_{ri} \sim \mathcal{N}(\theta_i, \sigma_i^2)$, for $r = 1, 2, \ldots$.
- Assumption 2: No CRN is used, i.e., Y_{ri} is independent of Y_{rj} for $i \neq j$.
- Assumption 3 (**indifference-zone**): The gap between the largest θ_i and the second largest θ_i is at least δ, *a value known to us*.
- Assumption 4 (known variance): σ_i^2 is known, for $i = 1, \ldots, k$.

- Bechhofer (1954) first developed a selection procedure, which can ensure the probability of correct selection (PCS):

$$\mathbb{P}\{\text{select the largest } \theta_i\} \geq 1 - \alpha, \tag{7}$$

under Assumptions 1-4, where α is a user specified value and $1 - \alpha > 1/k$.

- Bechhofer's Procedure
  1. Calculate a constant h, which satisfies
  $$\mathbb{P}\{Z_i \leq h,\ \ i = 1, 2, \ldots, k-1\} = 1 - \alpha, \tag{8}$$
  where $(Z_1, Z_2, \ldots, Z_{k-1})^\intercal$ has a multivariate normal distribution with means 0, variances 1, and common pairwise correlations $1/2$.
  2. For $i = 1, \ldots, k$, let
  $$n_i = \left\lceil \frac{2h^2\sigma_i^2}{\delta^2} \right\rceil. \tag{9}$$
  3. For $i = 1, \ldots, k$, run n_i replications for design i and calculate
  $$\bar{Y}_i = \frac{1}{n_i}\sum_{r=1}^{n_i} Y_{ri}.$$
  4. Select the design with the largest sample mean $\bar{Y}_i$ as the best.

Proof.

Without loss of generality, assume $\theta_k \geq \theta_{k-1} \geq \cdots \geq \theta_1$. Then Assumption 3 says, $\theta_k - \theta_{k-1} \geq \delta$, which implies that

$$\theta_k - \theta_i \geq \delta, \ i = 1, \ldots, k-1. \tag{10}$$

$$
\begin{aligned}
&\mathbb{P}\{\text{select } k\} = \mathbb{P}\{\bar{Y}_i - \bar{Y}_k < 0, \ i = 1, \ldots, k-1\} \\
=&\mathbb{P}\left\{\frac{\bar{Y}_i - \bar{Y}_k - (\theta_i - \theta_k)}{\sqrt{\sigma_k^2/n_k + \sigma_i^2/n_i}} < \frac{-(\theta_i - \theta_k)}{\sqrt{\sigma_k^2/n_k + \sigma_i^2/n_i}}, \ i = 1, \ldots, k-1\right\} \\
=&\mathbb{P}\left\{Z_i < \frac{\theta_k - \theta_i}{\sqrt{\sigma_k^2/n_k + \sigma_i^2/n_i}}, \ i = 1, \ldots, k-1\right\} \\
\geq&\mathbb{P}\left\{Z_i < \frac{\theta_k - \theta_i}{\sqrt{\sigma_k^2/(\frac{2h^2\sigma_k^2}{\delta^2}) + \sigma_i^2/(\frac{2h^2\sigma_i^2}{\delta^2})}}, \ i = 1, \ldots, k-1\right\} \quad \text{(due to (9))} \\
=&\mathbb{P}\left\{Z_i < \frac{\theta_k - \theta_i}{\delta/h}, \ i = 1, \ldots, k-1\right\} \\
\geq&\mathbb{P}\left\{Z_i < h, \ i = 1, \ldots, k-1\right\}. \quad \text{(due to (10))}
\end{aligned}
$$

Proof. (Cont'd)

Now we only need to check that $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_{k-1})^\intercal$ indeed has a multivariate normal distribution with means 0, variances 1, and common pairwise correlations $1/2$ (except for some rounding error).

Recall that

$$Z_i = \frac{\bar{Y}_i - \bar{Y}_k - (\theta_i - \theta_k)}{\sqrt{\sigma_k^2/n_k + \sigma_i^2/n_i}}, \quad i = 1, \ldots, k-1,$$

and $\boldsymbol{Y} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k)^\intercal$ is a k-variate normal random vector. So, $\boldsymbol{Z}$, as a linear combination of $\boldsymbol{Y}$, must be a $(k-1)$-variate normal random vector.

Besides, $\mathrm{Var}(Z_i) = \frac{\mathrm{Var}(\bar{Y}_i - \bar{Y}_k)}{\sigma_k^2/n_k + \sigma_i^2/n_i} = \frac{\sigma_k^2/n_k + \sigma_i^2/n_i}{\sigma_k^2/n_k + \sigma_i^2/n_i} = 1.$

Moreover, since $n_i = \left\lceil \frac{2h^2\sigma_i^2}{\delta^2} \right\rceil$ in (9), $\frac{\sigma_i^2}{n_i} = \frac{\delta^2}{2h^2}$ approximately, $i = 1, \ldots, k$.

For $i \neq j$, $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cov}\left(\frac{\bar{Y}_i - \bar{Y}_k}{\delta/h}, \frac{\bar{Y}_j - \bar{Y}_k}{\delta/h}\right) = \frac{\mathrm{Cov}(\bar{Y}_k, \bar{Y}_k)}{\delta^2/h^2} = \frac{\sigma_k^2/n_k}{\delta^2/h^2} = \frac{1}{2}.$

Hence, by (8) and (11), $\mathbb{P}\{\text{select } k\} \geq 1 - \alpha$. ■

- Assumption 3 (indifference-zone) can be **relaxed** by *softening* the selection target to probability of good selection (PGS):

$$\mathbb{P}\left\{\left|\text{selected } \theta_i - \max_{1\le i\le k} \theta_i\right| < \delta\right\} \ge 1-\alpha.$$

- Rinott (1978) proposed a procedure which can still guarantee the PCS in (7) while relaxing Assumption 4 (*known* variance), i.e., allowing *unknown* variances.
  - It requires an initial stage to estimate σ_i^2 by sample variance.
  - The proof is more complicated.

- Procedures like Bechhofer's or Rinott's are simple to implement, but the efficiency may be low.
  - The designed sample size (or, replication number), n_i, may be larger than necessary (too conservative).

- More sample efficient procedures should be in a sequential manner.
    - Take observations sequentially, i.e., one at a time.
    - Eliminate designs from continued sampling when it is statistically clear that they are inferior.
    - Simulation for a problem with a single dominant alternative may terminate very quickly.

- Paulson (1964) proposed fully sequential procedures, which can guarantee the PCS in (7), under Assumptions 1-3 and (a) *common known* variance or (b) *common unknown* variance.

- Suppose $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ and σ^2 is known (*common known* variance).
- Let $\bar{Y}_i(r)$ be the sample mean of the first r observations.
- Paulson's Procedure
  1. Let $0 < \lambda < \delta$ (a good choice is $\lambda = \delta/2$), and
  $$a = \ln\left(\frac{k-1}{\alpha}\right)\frac{\sigma^2}{\delta - \lambda}.$$
  Let $I = \{1, 2, \ldots, k\}$ and $r = 0$.
  2. Let $r \leftarrow r + 1$. Take one observation from each alternative in I and compute $\bar{Y}_i(r)$, $\forall i \in I$.
  3. Let $I^{\text{old}} = I$ and
  $$I = \left\{\ell \in I^{\text{old}} : \bar{Y}_\ell(r) \geq \max_{i \in I^{\text{old}}} \bar{Y}_i(r) - \max\{0, a/r - \lambda\}\right\}.$$
  If $|I| > 1$, then go to Step 2; otherwise, select the alternative left in I as the best.

- Kim and Nelson (2001) proposed a fully sequential procedure $\mathcal{KN}$, which extends Paulson's procedure, by allowing *unequal* variances and CRN.
- Commercial simulation software, Simio, implements the $\mathcal{KN}$ procedure of Kim and Nelson (2001) as an Add-In, to help user to select the best scenario.

- Ranking and Selection (R&S) problem was first introduced in the 1950s by the statistics community:
  - rank all alternatives
  - select a subset of alternatives
  - select the best alternative (attract the most attention)
- Existing procedures for R&S (selection of the best) problems:
  - frequentist
  - Bayesian

- Frequentist procedures typically aim to deliver the PCS or PGS; see Kim and Nelson (2006) for a review:
  - two-stage procedures: Bechhofer (1954), Rinott (1978)
  - sequential procedures: Paulson (1964), Kim and Nelson (2001), Hong (2006)

- Bayesian procedures often allocate samples to each alternative either to maximize the Bayesian posterior PCS or to minimize the expected opportunity cost; see Chen et al. (2015) for a review:
  - optimal computing budget allocation: Chen et al. (2000), He et al. (2007)
  - value of information: Chick and Inoue (2001), Chick et al. (2010)
  - knowledge gradient: Frazier et al. (2008), Frazier et al. (2009)
  - economics of selection procedures: Chick and Gans (2009), Chick and Frazier (2012)

- Emerging research problems that expend classical R&S from different perspectives; see Hong et al. (2021) for a review:
  - large-scale R&S using parallel computing
  - constrained R&S
  - multi-objective R&S
  - R&S with input uncertainty
  - R&S with covariates
- What if the number of candidate designs (feasible solutions) is huge, or countably infinite, or even uncountably infinite?
  - **Simulation Optimization (or called Optimization via Simulation)**

- R&S Problem vs Multi-Arm Bandit (MAB) Problem: