0405324: Stochastic System Simulation

Lecture 8: Validation and Verification of Simulation Models

Validation and verification

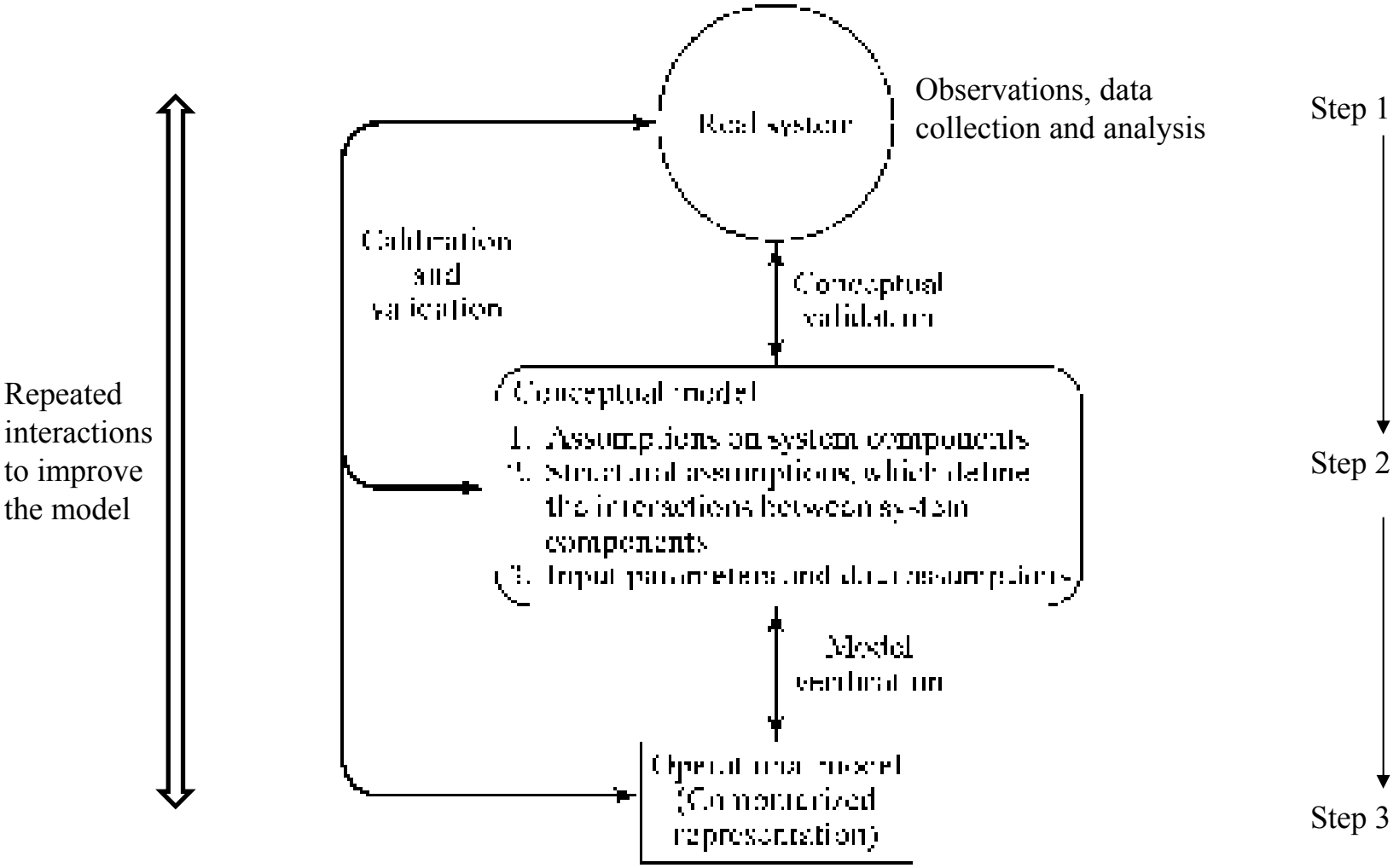
- Validation is the process of determining whether the conceptual model is an accurate representation of the actual system being analyzed:
  - Model that represent the system behavior accurately enough so that it can be used for experimentation
  - Increase credibility of the model so that it can be use by decision makers

  Validation deals with building the right model.

- Verification is the process of determining whether a simulation computer program works as intended
  - Making sure that the conceptual model is implemented correctly (input parameters, logical structure) in the computer program (Arena)
  - Debugging the computer program

  Verification deals with building the model right.

Modeling-Building, Verification & Validation



Techniques for Verification of Simulation Models

- **Use good programming practice:**
  - Write and debug the computer program in modules or subprograms.
  - In general, it is always better to start with a "moderately detailed" model, and later embellish, if needed.
- **Use "structured walk-through":**
  - Have more than one person to read the computer program.

Techniques for Verification of Simulation Models

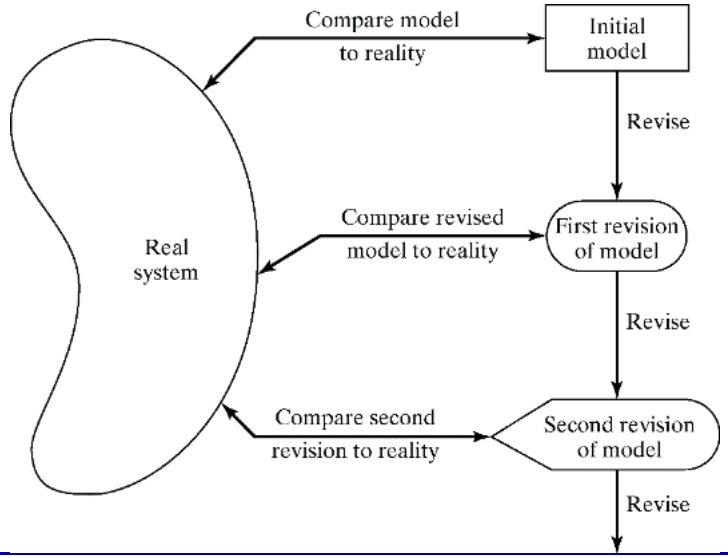
- **Check simulation output for reasonableness:**
  - Run the simulation model for a variety of input scenarios and check to see if the output is reasonable.
  - In some instances, certain measures of performance can be computed exactly and used for comparison.
- **Use a "trace":**
  - The analyst may use a trace to print out some intermediate results (or for the different scenarios) and compare them with hand calculations to see if the program is operating as intended.
- **Animate:**
  - Using animation, the users see dynamic displays (moving pictures) of the simulated system.
  - Since the users are familiar with the real system, they can detect programming and conceptual errors.

Techniques for Verification of Simulation Models

- **Compare final simulation output with analytical results:**
  - May verify the simulation response by running a simplified version of the simulation program with a known analytical result. If the results of the simulation do not deviate significantly from the known mean response, the true distributions can then be used.
  - For example, for a queuing simulation model, queuing theory can be used to estimate steady state responses (e.g., mean time in queue, average utilization). These formulas, however, assume exponential interarrival and service times with n servers (M/M/n).

Calibration and Validation

- **Validation**: the overall process of comparing the model and its behavior to the real system.
- **Calibration**: the iterative process of comparing the model to the real system, making adjustments, comparing again, and so on (until the model output is acceptable).
- No model is ever a perfect representation of the system
  - The modeler must weigh the possible, but not guaranteed, increase in model accuracy versus the cost of increased validation effort.



Techniques for Validation of Simulation Models

- **Comparison of the model with the real system:**
  - **Subjective tests**: people, knowledgeable about the system, make judgments about the model and its outputs.
  - **Objective tests**: requires data on the system's behavior, and the corresponding data produced by the model → statistical tests to compare the two data sets
- **A three-step approach for developing a valid and credible model:**
  - Build a model that has high face validity.
  - Validate model assumptions.
  - Compare the model input-output transformations with the real system's data.

Techniques for Validation of Simulation Models: High Face Validity [Calibration & Validation]

- The objective of this step is to develop a model that, on the surface, seems reasonable to people who are familiar with the system under study.
- This step can be achieved through discussions with system experts, observing the system, or the use of intuition.
- It is important for the modeler to interact with the client on a regular basis throughout the process.
  - Users can identify model's deficiencies
  - It increases model's perceived credibility for managers
- Sensitivity analysis can also be used to check a model's face validity. For example: In most queueing systems, if the arrival rate of customers were to increase, it would be expected that server utilization, queue length and delays would tend to increase.

Techniques for Validation of Simulation Models: Model assumptions' Validity [Calibration & Validation]

- **General classes of model assumptions:**
  - Structural assumptions: how the system operates.
  - Data assumptions: reliability of data and its statistical analysis.
- **Bank example: customer queuing and service facility in a bank.**
  - Structural assumptions, e.g., customer waiting in one line versus many lines, served FCFS versus priority.
    - Verify structural assumptions by actual observations and discussions with managers and users.
  - Data assumptions, e.g., interarrival time of customers, service times for commercial accounts.
    - Verify data reliability with bank managers.
    - Data can also be verified by using correlation (between data collected in different periods) and goodness of fit tests (distributions)

Techniques for Validation of Simulation Models: Validate input-output transformation [Calibration & Validation]

- **Goal: Validate the model's ability to predict future behavior of the system**
  - When the model input data match the real inputs and when a policy implemented in the system is implemented in the model
  - This is the only objective test of the model.
  - The structure of the model should be accurate enough to make good predictions for the range of input data sets of interest (if one or more input change, the model should predict the change in the real output).
- **The model is seen as an input-output transformation: this correspondence is being validated.**

- **One possible approach: use historical data that have been reserved for final validation purposes only (other than the ones used for model development and calibration).**
- **Condition:** the system, or a version of it at least, exists (with input and output data).
- **Criteria:** use the main responses of interest to validate the model (e.g., in a production system: the throughput).

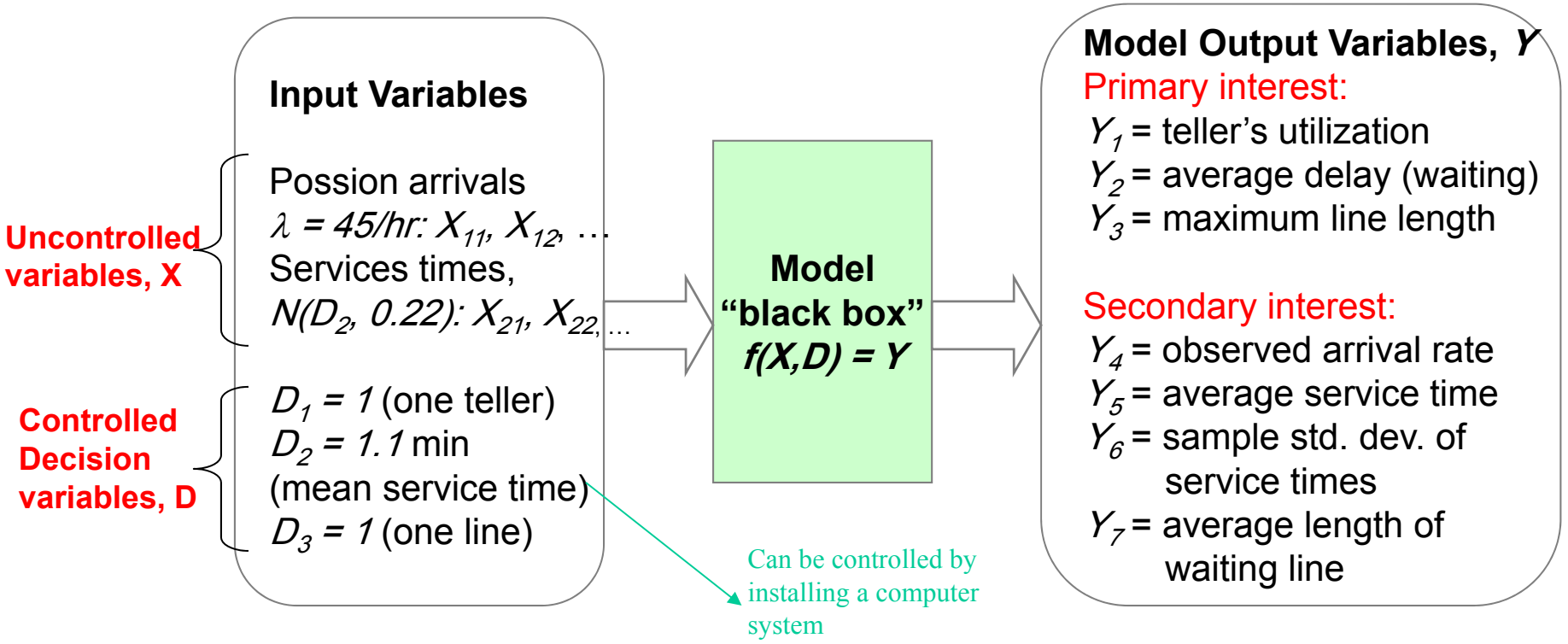
Bank Example [Validate I-O Transformation]

- **Example: One drive-in window serviced by one teller, only one or two transactions are allowed.**
  - Data collection: 90 customers during 11 am to 1 pm (consultation with the management → this is a typical rush hour).
    - Observed service times $\{S_i, i = 1,2, \ldots, 90\}$.
    - Observed interarrival times $\{A_i, i = 1,2, \ldots, 90\}$.
  - Data analysis led to the conclusion that:
    - Interarrival times: exponentially distributed with rate $\lambda = 45$ customer/h
    - Service times: $N(1.1, 0.2)$

The Black Box

[Bank Example: Validate I-O Transformation]

- **A model was developed in close consultation with bank management and employees (verified)**
- **Model assumptions were validated (as discussed earlier)**
- **Resulting model is now viewed as a "black box" transforming inputs into outputs:**



Comparison with Real System Data

[Bank Example: Validate I-O Transformation]

- **Real system data (historical record) are necessary for validation.**
  - System responses should have been collected during the same time period (from *11*am to *1*pm on the same Friday) in order to be compared with the simulation outputs.
- **Compare the average delay from the model Y_2 with the actual delay Z_2 (average observed delay on Friday 11am-1pm):**
  - Suppose that the average delay observed, $Z_2 = 4.3$ minutes, consider this to be the true mean value $\mu_0 = 4.3$ (for the purpose of validation).
  - When the model is run with generated random variates X_{1n} and X_{2n}, Y_2 should be close to Z_2.
  - Let's see how the modeler would check this consistency (between Y_2 and Z_2)
  - If Y_2 is the simulation output, and $\mu = E(Y_2)$,
  - Six statistically independent replications of the model, each of *2*-hour duration, are run. The results are provided in slide 15.

Comparison with Real System Data

[Bank Example: Validate I-O Transformation]

Replication	Y_4 *(Arrivals/Hour)*	Y_5 *(Minutes)*	Y_2 = *Average Delay (Minutes)*
1	51	1.07	2.79
2	40	1.12	1.12
3	45.5	1.06	2.24
4	50.5	1.10	3.45
5	53	1.09	3.13
6	49	1.07	2.38

Example-

Output data on average delay from the simulation model for six replications are given in the above shown Table. Do you think that the simulation model is a valid one if the actual mean delay is 4.3 minutes (use 5% level of significance)

Hypothesis Testing

[Bank Example: Validate I-O Transformation]

- **The hypothesis testing evaluates whether the simulation and the real system are *the same* with respect to some output performance measure or measures.**
- **Compare the average delay from the model Y_2 with the actual delay Z_2 (continued):**
  - Null hypothesis testing: evaluate whether the simulation and the real system are *the same* (w.r.t. output measures):

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

    - If H_0 is not rejected, then, there is no reason to consider the model invalid
    - If H_0 is rejected, the current version of the model is rejected, and the modeler needs to improve the model

Hypothesis Testing
[Bank Example: Validate I-O Transformation]

- Solution: Conduct the *t* test:
  - Hypothesis:

$$H_0: \mu = 4.3 \text{ minutes}$$
$$H_1: \mu \neq 4.3 \text{ minutes}$$

  - Compute test statistics:

$$\bar{Y}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_{2i} = 2.51 \text{ minutes} \qquad S = \sqrt{\frac{\sum_{i=1}^{n}(Y_{2i} - \bar{Y}_2)^2}{n-1}} = 0.82 \text{ minutes}$$

$$t_0 = \frac{\bar{Y}_2 - \mu_0}{S/\sqrt{n}} = \frac{2.51 - 4.3}{0.82/\sqrt{6}} = -5.34 \; < \; t_{0.025,5}(= -2.571), (t_{\alpha/2,n-1} : \text{two-sided test})$$

  - Decision: reject H_0. Conclude that the model is inadequate.
  - Check: the assumptions justifying a *t* test, that the observations (Y_{2i}) are normally and independently distributed.

Hypothesis Testing

[Bank Example: Validate I-O Transformation]

- **Similarly, compare the model output with the observed output for other measures:**
  $Y_4 \leftrightarrow Z_4$, $Y_5 \leftrightarrow Z_5$, **and** $Y_6 \leftrightarrow Z_6$

Type I and type II Errors

[Validate I-O Transformation]

Table Types of Error in Model Validation

Statistical Terminology	*Modeling Terminology*	*Associated Risk*
Type I: rejecting H_0 when H_0 is true	Rejecting a valid model	α
Type II: failure to reject H_0 when H_1 is true	Failure to reject an invalid model	β

- **Type I error (α):**
  - Error of rejecting a valid model.
  - Controlled by specifying a small level of significance α (say, 0.1, 0.05, or 0.01).
- **Type II error (β):**
  - Error of accepting a model as valid when it is invalid.
  - *Controlled by specifying critical difference, δ (between E(Y) and μ_0) and find the n.*
- **A type II error is more serious than type I error, and thus, it is important to design a simulation experiment to control this risk.**

Type II Error

[Validate I-O Transformation]

- **For validation, the power of the test is:**
  - Probability [detecting an invalid model] = $1 - \beta$
  - The test power is the probability of detecting a departure from H_0, when in fact, such a departure exists→ the probability of detecting an invalid simulation model.
  - β = P(Type II error) = P(failing to reject H_0/H_1 is true)
- **The failure to reject H_0 is a weak conclusion unless the test power (1- β) is large (close to 1) → β is small.**
  - Value of β depends on:
    - Sample size, n
    - The true difference, δ, between $E(Y)$ and μ_0: $\delta = \frac{|E(Y) - \mu_0|}{\sigma_0}$

      where, σ_0 (population standard deviation) that can be estimated by S, with
- **In general, the best approach to control β error is:**
  - For a specified value of the critical difference, δ.
  - Choose a sample size, n, by the operating characteristics curve (OC curve), which are curves of β versus δ.

Type II Error: OC curve

[Validate I-O Transformation]

Table Operating Characteristic Curves for The Two-Sided t Test for Different Values of Sample Size n

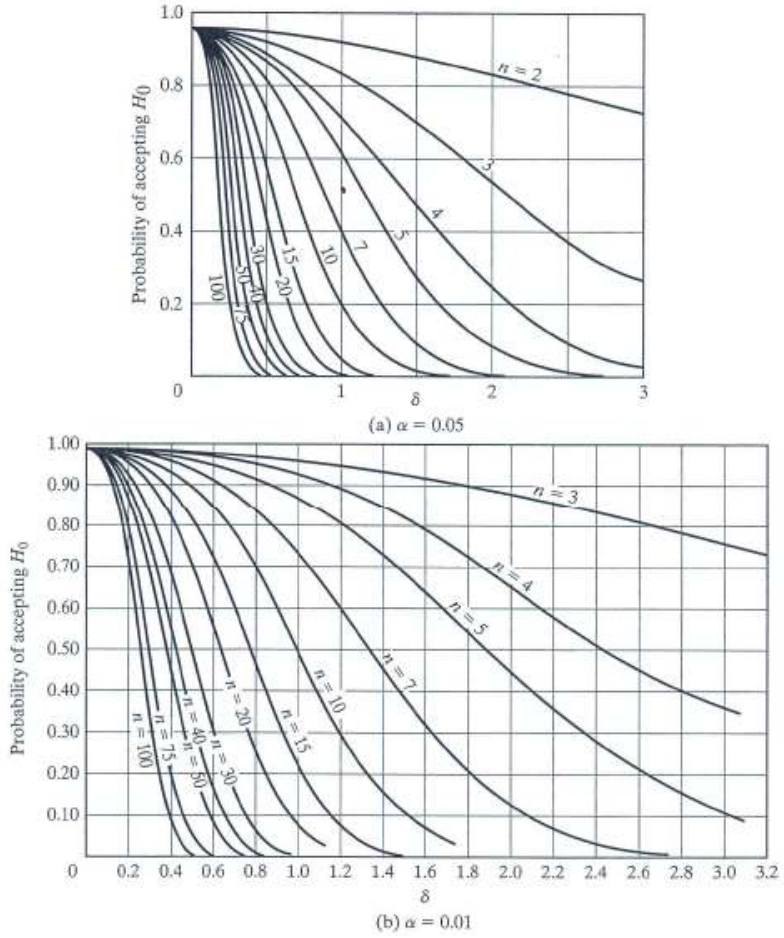
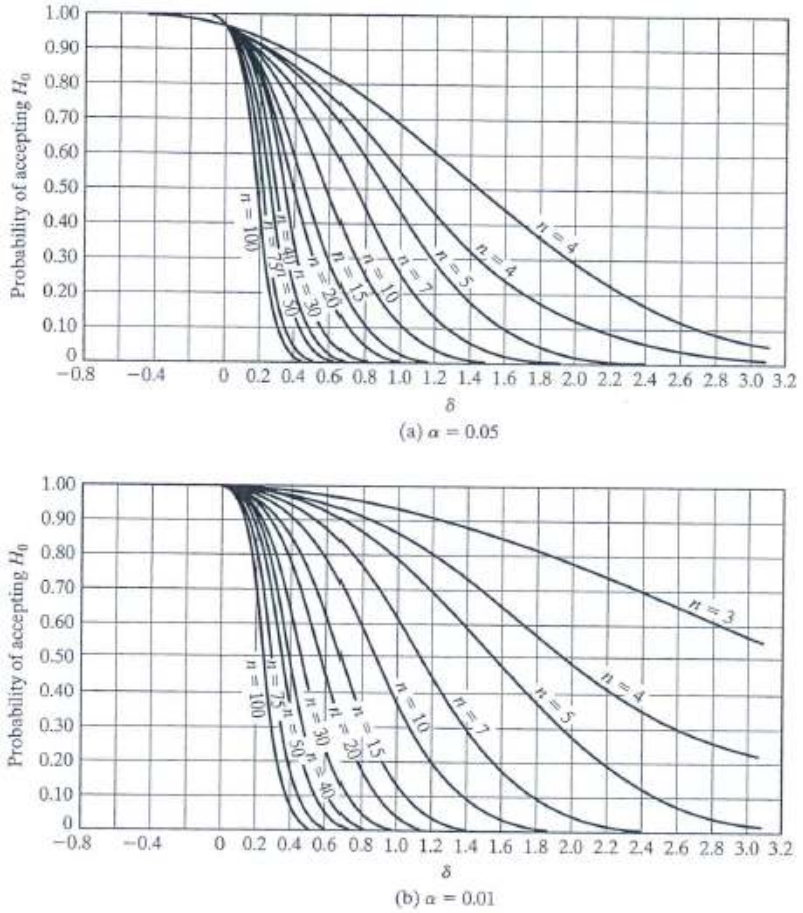


Table Operating Characteristic Curves for the One-Sided t Test for Different Values of Sample Size n



*Check blackboard

Type II Error: OC curve

[Validate I-O Transformation]

Example-

After rejecting H_0, in the example in slide 17, the modeler has discovered some mistakes in the model, and revised it. New 6 replications have been run. The output data on average delay from these six replications are given in the following Table. Using 5% level of significance for a two-sided test

(i) Find the probability of accepting an invalid model, if the true mean delay of the model differed from the average delay in the system by 1 minute,

(ii) Suppose that the modeler would like to reject invalid model with probability at least 0.90 (i.e., $1-\beta \geq 0.9 \rightarrow \beta \leq 0.1$), what should the replicate size for the average delay by 1 minute?

Replication	Y_4 *(Arrivals/Hour)*	Y_5 *(Minutes)*	Y_2 = *Average Delay (Minutes)*
1	51	1.07	5.37
2	40	1.11	1.98
3	45.5	1.06	5.29
4	50.5	1.09	3.82
5	53	1.08	6.74
6	49	1.08	5.49

Type II Error: OC curve

[Validate I-O Transformation]

Solution- Based on the data given in the table,

$n = 6,\ \alpha = 0.05$

$\bar{Y}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_{2i} = 4.78$ minutes $\qquad S = \sqrt{\frac{\sum_{i=1}^{n}(Y_{2i} - \bar{Y}_2)^2}{n-1}} = 1.66$ minutes

(i) $\hat{\delta} = \frac{|E(Y) - \mu_0|}{S} = \frac{1}{1.66} = 0.60$

At $n = 6, \hat{\delta} = 0.06$ from OC curve, $\beta = 0.75$– quite high

(ii) Given, Power = 1- $\beta \geq 0.90$, i.e. $\beta \leq 0.10$,

At $\beta \leq 0.10$, and $\hat{\delta} = 0.6$, from OC curve, $n = 30$ is required to achieve a power of 0.90. Thus the additional replicate needed = 30 – 6 = 24.

Confidence Interval Testing

[Validate I-O Transformation]

- **Confidence interval testing: evaluate whether the simulation and the real system are *close enough.***
- **Assume that there is a known output performance measure for the existing system, μ_0, and unknown performance measure of the simulation, μ, that we hope is close. The hypothesis testing tests whether $\mu = \mu_0$; the C.I. formulation tries to bound the difference $|\mu - \mu_0|$ to see whether it is $\leq \varepsilon$, a difference that is small enough to allow valid decisions to be based on the simulation. The value of ε is set by the analyst.**
- **If Y is the simulation output, and $\mu = E(Y)$, the confidence interval (C.I.) for μ is:**

$$\bar{Y} \pm t_{\alpha/2,n-1} S/\sqrt{n}$$

Confidence Interval Testing

[Validate I-O Transformation]

- **Validating the model:**
  - Suppose the C.I. does not contain μ_0 (see Figure (a), below):
    - If the best-case error is $> \varepsilon$, model needs to be refined.
    - If best-case error is $\leq \varepsilon$, additional replications are necessary (to shrink the CI).
    - If the worst-case error is $\leq \varepsilon$, accept the model.
  - Suppose the C.I. contains μ_0 (see Figure (b), below):
    - If either the best-case or worst-case error is $> \varepsilon$, additional replications are necessary (to shrink the CI).
    - If the worst-case error is $\leq \varepsilon$, accept the model.

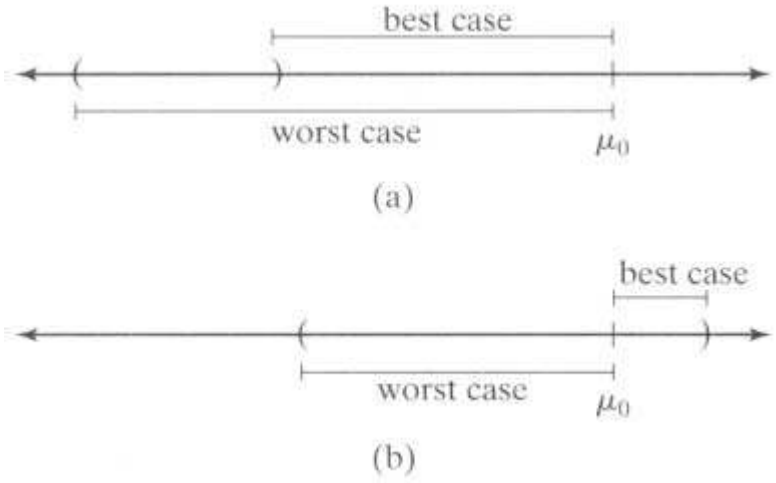


Figure. Validation of the I-O transformation:

(a) when true value falls outside,

(b) (b) when true value falls inside, the C.I.

Confidence Interval Testing

[Validate I-O Transformation]

Example: From the data on slides # 15 and 17, $\mu_0 = 4.3$, and "close enough" is $\varepsilon = 1$ minute (slide 22) of expected customer delay.

- A *95%* confidence interval, based on the *6* replications is *[1.65, 3.37]* because:

$$\bar{Y} \pm t_{0.025,5} S / \sqrt{n}$$

$$2.51 \pm 2.571(0.82 / \sqrt{6})$$

- μ_0 = 4.3 falls outside the confidence interval, the best case **|3.37 – 4.3| = 0.93 < 1**, but the worst case **|1.65 – 4.3| = 2.65 > 1**, therefore, additional replications are needed to reach a decision.
- How to find out the number of additional replications needed??

Half Width, Number of Replications (cont'd.)

- **Set half-width = *h*, solve for** $n \geq \left(\frac{t_{\alpha/2,R-1} s_0}{h} \right)^2$ **(*n* is replicate size, *R*)**
- **Not really solved for *n* (*t*, *s* depend on *n*)**
- **Approximation:**
  - Replace *t* by *z*, corresponding normal critical value
  - Pretend that current *s* will hold for larger samples
  - Get $R \geq \left(\frac{z_{\alpha/2} s_0}{h} \right)^2$

s_0 = sample standard deviation from "initial" number n_0 of replications

- **Easier but different approximation:**

$$n \cong n_0 \frac{h_0^2}{h^2}$$

h_0 = half width from "initial" number n_0 of replications

Note: See Lecture 7, Slide # 21-26.

Using a Turing Test

[Validate I-O Transformation]

- **The Turing test is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human**
- **Use Turing test in addition to statistical test, or when no statistical test is readily applicable.**
- **Utilize persons' knowledge about the system.**
- **For example:**
  - Present 10 system performance reports to a manager of the system. Five of them are from the real system and the rest are "fake" reports based on simulation output data.
  - If the person identifies a substantial number of the fake reports, interview the person to get information for model improvement.
  - If the person cannot distinguish between fake and real reports with consistency, conclude that the test gives no evidence of model inadequacy.

See Tutorial-4 (Lecture-8)

End of lecture