Output Analysis

CS1538: Introduction to Simulations

Output Analysis for a Single Model

- Since most simulations are stochastic in nature, their output can vary from run to run due to random chance
  - The results from any single run may not be useful

- We typically need to analyze our results over many runs
  - The values may have great differences
  - No one run would necessarily represent the "correct" result
  - We need to perform statistical analysis on these results in some way

Model output types

- The analysis is affected by the type of outputs

- They generally fall into two categories of behaviors for a stochastic process:
  1) Transient behavior
     - Indicated by a simulation with a specific termination event (ex: runs for X minutes, or runs until C customers have been processed, or runs until inventory is exhausted etc.)
  2) Steady-state behavior
     - Indicated by a simulation that runs over a very long period of simulated time, or with no stated stop event

Transient vs. Steady State

- Mathematically, their output distributions have different characteristics
  - Consider output $Y_1, Y_2, \ldots$ for a simulation with initial conditions *I*
  - And distribution $F_r(y \mid I) = P(Y_r \leq y \mid I)$ for $r = 1, 2, \ldots$
    - In words, F gives the probability distribution for each output conditional upon the initial conditions
  - For a stochastic process with transient behavior
    - Each F_r will be different for different r and for different I
    - In other words the output values vary from each other and are different for different initial conditions
  - For a stochastic process with steady state behavior
    - Eventually a point is reached such that

      $F_r(y \mid I) \rightarrow F(y)$
    - Or, the output converges to a distribution that is independent of r and I

Transient Behavior Processes

- Consider an output value of interest
  - Ex: Queue length, Q
  - Ex: Wait time in queue, W^Q
- Within a single run of a simulation, the values will autocorrelate, and thus will not be independent
  - *Why might that be?*
  - Because they are not IID (Independent & Identically Distributed), we cannot do "classical" statistical analysis on these values *within* a single simulation run
  - However, the results should be independent *between* runs, as long as different random number streams were used.

Transient Behavior Processes

- For example, suppose W_{cr}^{Q} represents the wait time for customer c during run r of the simulation
  - W_{cr}^{Q} for the same r and c = 1, 2, … are *not* independent
  - W_{cr}^{Q} for the same c and r = 1, 2, … *are* independent, and can be analyzed as such
  - If we get the average wait for each run, we can also analyze that value across the runs
- To analyze the values, we'd like to establish some confidence about the accuracy / validity of our results

Confidence Intervals

- Consider result Y_r for r = 1, 2, …, n that we obtain for the n runs of our simulation
  - We refer to the (ground truth) mean of the result as μ and the standard deviation σ
    - However, we don't know what they are (since we only have some observed data)
  - We'd like to know whether the computed mean of the observed results, $\bar{Y}$, is close to the actual mean, μ, of the output distribution
- A confidence interval will tell us whether the actual mean is within a certain range with a certain probability
  - Ex: $\mu = 5.2 \pm 0.32$ with confidence 90%

    This says that there is a 90% chance that the actual mean of our output distribution is between 4.88 and 5.52

Confidence Intervals

- Desirable properties of the confidence interval:
  - The confidence to be as high as possible (close to 1)
  - The interval to be as narrow as possible (width close to 0)
- The actual confidence and interval values are dependent upon a few factors, including:
  - The variation in the data produced by the simulation
  - The number of simulation runs performed
- There are two methods of determining confidence intervals
  1. Given **n runs and a confidence probability**, what is our confidence interval?
  2. Given **a desired confidence interval**, how many runs are necessary to achieve that interval?

Confidence Intervals

- First, we determine the point estimate, $\bar{Y}$, for our distribution
  - This is simply the mean of the sample points

$$\bar{Y} = \frac{\sum_{j=1}^{n} Y_j}{n}$$

- We next need to determine the sample variance, S^2

$$S^2 = \frac{\sum_{j=1}^{n} \left(Y_j - \bar{Y}\right)^2}{n-1}$$

Confidence Intervals

- This gives us an estimate of the actual variance, σ^2, of $\bar{Y}$ through the following:

$$\bar{\sigma}^2(\bar{Y}) = \frac{S^2}{n}$$

- Taking the square root of this gives us the standard error of the sample mean

$$\bar{\sigma}(\bar{Y}) = \sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}}$$

- This value helps us to determine how accurate our estimate of the mean, $\bar{Y}$, is.

Confidence Intervals

- As long as our estimators are not too biased, the random variable is approximately t-distributed with n-1 degrees of freedom

$$1-\alpha \approx P\left(-t_{\alpha/2,n-1} \le \frac{\bar{Y}-\mu}{\sqrt{S^2/n}} \le t_{\alpha/2,n-1}\right) =$$

$$P\left(-t_{\alpha/2,n-1}\sqrt{S^2/n} \le \bar{Y}-\mu \le t_{\alpha/2,n-1}\sqrt{S^2/n}\right) =$$

$$P\left(\bar{Y}-t_{\alpha/2,n-1}\sqrt{S^2/n} \le \mu \le \bar{Y}+t_{\alpha/2,n-1}\sqrt{S^2/n}\right)$$

- Hence, we are (1-α)% confidence that μ is within:
  - $\bar{Y}-t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S^2}{n}}$ and $\bar{Y}+t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S^2}{n}}$

Confidence Intervals

- Using the formula, we get the interval around our point estimate of the mean
- It says that the actual mean will fall in the interval with a $(1 - \alpha)$ probability
  - We look up the α value in the t-distribution table and plug in the other numbers
  - We can also think of it visually, looking at a t-distribution curve
    - It is actually quite similar to a normal curve and is also symmetric
    - The α value is divided into $\alpha/2$ on either extreme of the curve
  - The t-distribution approaches a standard normal distribution as degrees of freedom $\rightarrow\infty$
    - If n is large enough we can substitute z for t

Exercise

- A call center takes questions between 8am and 4pm. The results of a simulation of the call center is shown below.

Run	Avg Wait Time (in minutes)	Avg # of people on hold
1	0.88	0.68
2	5.04	4.18
3	4.13	3.26
4	0.52	0.34

- What is the sampled average wait time?
- Find the 95% confidence interval around the sample average wait time.

Confidence Intervals

- Suppose we'd like our sample mean to be within 2ε of the actual mean, with a stated probability of $(1-\alpha)$, i.e.:
  - $P(|\bar{Y} - \mu| < \epsilon) \geq 1 - \alpha$

- We will do this in a multistep process:
  - Choose an n_0 as an initial number of runs to try. From n_0 we calculate an S_0 (initial standard deviation)
  - We know that the precision increases (and ε decreases) with increased n

Confidence Intervals

- From our previous derivations, we know that the half-length of a confidence interval is
  - half-length = $\frac{t_{\alpha/2,n-1} S_0}{\sqrt{n}} \leq \varepsilon$
  - for confidence probability 1 - α
- Which we would like to solve for n such that

$$n \geq \left(\frac{t_{\alpha/2,n-1} S_0}{\varepsilon} \right)^2$$

- However, since we don't know n yet, we can't look up $t_{\alpha/2,\, n-1}$
  - But we do know that the t-distribution approaches the standard normal distribution as n $\rightarrow \infty$
  - If we substitute $z_{\alpha/2}$ we can get a ballpark value for n
    - Last row of Student's T Table

Confidence Intervals

- This leads us to $$n \geq \left(\frac{z_{\alpha/2} S_0}{\varepsilon} \right)^2$$
- However, this may not be exactly correct, since we substituted z for t in our formula
  - Since $t_n > z$, the value we calculate for n may be a bit small
    - We can make this an iterative process – updating n and testing it until we get the desired precision

Exercise

- Same setup as before.

Run	Avg Wait Time (in minutes)	Avg # of people on hold
1	0.88	0.68
2	5.04	4.18
3	4.13	3.26
4	0.52	0.34

- Suppose we want the confidence interval to be within 0.5 minutes with 95% confidence. How many *more* runs do we have to do?

Quantiles and Percentiles

- Sometimes, instead of the average value, we are more interested in the probability that the result for a given run will be ≤ some value
  - In other words, we want to specify some probability p such that

    $\Pr(Y \leq \theta) = p$

    for some value θ
  - We say the value θ is the p^{th} quantile (as a fraction) or percentile (as a percentage) of Y
    - *Example*: Consider a simulation of a grocery checkout line. Suppose we run over several (independent) days, keeping track of the average time in the system for each customer
      - Determine the value for θ, such that the average time a customer spends in the system is ≤ θ on 80% of the days

Quantiles and Percentiles

- We typically determine these in the following way:
  - Determine the desired probability
  - Find the value that satisfies that probability
- More specifically:
  - Calculate the point estimate, by taking the appropriate proportion of the sample points
    - Ex: If we have 1000 sample points and we want the 80th percentile, we estimate the 80th percentile as the 0.8(1000)th point (sorted)
  - We then calculate the (1 - α) confidence interval using formulas 11.18 in the Banks et al. text

$$p_l = p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n-1}} \qquad p_u = p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n-1}}$$

Quantiles and Percentiles

- These give us the range of probabilities that we then convert to quantiles
  - Ex: The lower bound, θ_l, is the $p_l(1000)^{th}$ point in our sample (sorted) and the upper bound, θ_u, is the $p_u(1000)^{th}$ point in our sample (sorted)
- Ex: Thus the 80^{th} percentile of our sample will be between θ_l and θ_u with probability $(1 - \alpha)$
- This gets a bit confusing because we are talking about probabilities at two different levels
  - The quantile itself gives a probability
  - The confidence interval also gives a probability

Percentiles Example

- Consider a simulation of a grocery checkout line. Suppose we run over several (independent) days, keeping track of the average time in the system for each customer.
- Determine the value for θ (with 95% confidence), such that the average time a customer spends in the system is $\leq \theta$ on 80% of the days

Average time for each run (minutes)

5.4	10.6
3.7	4.5
8.3	8.5
6.1	9.3
2.2	6.4

Comparing Two Alternative Designs

- If we have the same output variable of interest for both designs, we'd like to compare them
  - Are the differences between them statistically significant?
  - A common approach is to look at the difference of the values
    - Form a confidence interval around the difference
    - If the difference is large and the confidence interval around it is tight and with high confidence, we say that the difference is statistically significant
    - We need to judge separately whether the difference is **practically** significant

Comparing Two Alternative Designs

- What if the difference between the two alternatives led to an observed difference within the following confidence interval:
  - $-4.5 \leq \bar{Y}_1 - \bar{Y}_2 \leq 3.6$ with 95% confidence
- What if the difference has the following confidence interval instead:
  - $3.0 \leq \bar{Y}_1 - \bar{Y}_2 \leq 4.3$ with 95% confidence

Comparing Two Alternative Designs

- Correlated Sampling
  - Identical random data streams are used for the different simulation variations
    - i.e. the input data is identical for both versions (not just the same distribution, but the same exact data)
    - Will always have the same number of runs in this case
  - Correlated sampling tends to reduce variance in the difference of the results, thereby giving better confidence intervals for the same number of runs
    - However, it is not always possible or easy to do

- Independent Sampling
  - Separate random data streams are used for the different simulation variations
  - We may or may not have the same number of runs for each $\bar{Y}_i$

Correlated Sampling

- Also called paired-t approach
- In this case, both versions use the identical random data for the identical number of runs
  - Thus, the output from run X_i for each version is not independent
- So we can process the difference of the versions $Y_{r1} - Y_{r2}$ as a new, separate random variable for each run r
- Then the difference result can be analyzed in the same way as single result data
- An advantage of this approach is that it reduces the variance (which, in turn reduces the width of the confidence interval)

Exercise

- We ran simulations to compare two methods of taking customer orders. Both models received identical input data for all 10 trials. To the right are the average wait time for each method.
- Based on the outcome, can we be 95% confident that the methods are different?

Trial	Method A	Method B
1	12.3	12.0
2	12.0	12.3
3	12.0	12.5
4	13.0	12.0
5	13.0	13.1
6	12.5	12.4
7	11.3	10.3
8	11.8	11.3
9	11.5	11.6
10	11.0	11.5

Independent Sampling

- Trials from one simulation variant have no pair/connection to trials from another simulation variant

- Two possibilities:
  - Equal Variances
  - Unequal Variances
  - *Why didn't we have this question with paired sampling?*

Independent Sampling with Equal Variances

- Also called two-sample-t approach
- Method 1 and Method 2 have the same variance $\sigma_1 = \sigma_2$
  - Method 1 is run for R_1 trials
  - Method 2 is run for R_2 trials

- What is the confidence interval for the difference of outcomes of the two methods?
  - $(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2,v}$ [s.e.$(\bar{Y}_1 - \bar{Y}_2)$] with prob. $(1-\alpha)$
    - To calculate this we need to determine the standard error (s.e.) for the difference of the two means, and the degrees of freedom (v)

- Degrees of freedom is just sum of df of both methods:
  - $(R_1 - 1) + (R_2 - 1) = R_1 + R_2 - 2$
- Calculating the standard error is a little more complicated
  - Since $R_1 \neq R_2$, the contributions from the two methods' sample variances may not be equal

Independent Sampling with Equal Variances

- Weighted average of the two sample variances:

$$S_p^2 = \frac{(R_1 - 1)S_1^2 + (R_2 - 1)S_2^2}{R_1 + R_2 - 2}$$

  - Note that if $R_1 = R_2$, the pooled sample variance is simplified to just the sum of the two sample variances divided by 2.

- The standard error is then:

$$s.e.(\bar{Y}_1 - \bar{Y}_2) = S_p \sqrt{\frac{1}{R_1} + \frac{1}{R_2}}$$

Independent Sampling with Equal Variances

- Can generally be used in two situations:
  - Variances for Y_1 and Y_2 *are the same* but the number of runs of each *may or may not* be the same
  - The number of runs for Y_1 and Y_2 *are the same* and the variances *may or may not* be the same

Independent Sampling with Unequal Variances

- Also called modified two-sample-t approach
- For when variances for Y_1 and Y_2 are not equal
- Standard Error

$$s.e.(\overline{Y_1} - \overline{Y_2}) = \sqrt{\text{var}(\overline{Y_1}) + \text{var}(\overline{Y_2})} = \sqrt{\frac{S_1^{\,2}}{R_1} + \frac{S_2^{\,2}}{R_2}}$$

- Degrees of Freedom

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

- Confidence Interval
  - $(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2,\nu}$ [s.e.$(\bar{Y}_1 - \bar{Y}_2)$] with prob. $(1-\alpha)$

Equality of Variances?

- Homoscedasticity – homogeneity of variances
- Test for homoscedasticity
  - Bartlett's test – assumes normality of data
  - Levene's test – less sensitive to departures from normality
  - Brown-Forsythe test – even less sensitive (tests on median, not mean)

- H_0: variances are equal
- H_a: variances are not equal

Conclusions

- If identical random data and identical runs can be used for both design versions, it should be
  - A positive correlation between the versions reduces the variance, thereby improving the confidence interval
  - Note that caution must be taken to ensure that the data used is actually identical
    - So for each data stream, value X_i in version 1 must be used for the same purpose in version 2

Conclusions

- In some circumstances, using identical data may not be possible
  - One version may have different amounts of input from the other, or different input parameters
  - In these cases, we need to use independent sampling, as discussed previously
  - If possible, we'd like the number of runs to be the same
- Generally speaking, we prefer to get results that will have the least variance, thereby giving the best confidence intervals
  - However, we can always improve our confidence intervals for any system by increasing the runs

Which Approach to Use?

- Have two methods or versions of the system.

- Correlated (Paired) Sampling
  - Same number of runs for both methods
  - For both methods, run *i* has the same random number stream

- Independent Sampling
  - Equal Variances vs. Unequal Variances
    - Test for equality of variances using Homoscedasticity test
    - Could always assume unequal variances, but not as powerful as equal variance version

Steady State Analysis

- If our simulation is evaluating steady state behavior, we use somewhat different evaluation techniques
- Recall that for a given measure θ, if we are determining a long run value, then

$$\theta = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} Y_i$$

  - For some variable of interest Y, where n is the number of observations of Y
  - Where θ is independent of any initial conditions of the simulation

Steady State Analysis

- However, it may not be obvious **how large n has to be** until we approach the long run value
- Initial conditions CAN affect the value of n needed to get to a steady state
  - These introduce a bias into the data that can take a lot of time to dissipate

Steady State Analysis

- Some ways to reduce this initial bias:
  - Set initial conditions intelligently
  - Run the simulation for a while, then start collecting data

Steady State Analysis

1) Set the initial conditions of the simulation in an intelligent way, as close to the steady state values as possible
   - These conditions could be obtained through observation of the real system
   - They could also be approximated using analytical techniques (Markov analysis, as discussed in Chapter 6 of Banks et al)
     - Determine long run values and use these as the initial conditions for our simulation
     - We likely cannot model the exact system but we can get something close enough to reduce the initial bias

Steady State Analysis

2) Run the simulation for a while without tabulating any data
   - By the time data begins to be collected, the conditions should be close to the steady state

   - How can we determine how long to wait before tabulating data?
   - In other words, if we divide our run into two segments, T_0 and T_E, how long should we make T_0 and T_E be?
     - This is most likely done via experimentation
     - The text discusses how to break up runs into batches and calculate ensemble averages in order to accomplish this
   - Idea:
     - Let's say we want to do R runs of our simulation
     - Let's say each run will be for time T

Steady State Analysis

- Break up T into k subdivisions (batches)
- Given variable of interest, Y, calculate the average value of Y for each batch across all runs to get the ensemble averages

$$\overline{Y}_{.j} = \frac{1}{R}\sum_{r=1}^{R} Y_{rj} \qquad \text{for } 1 <= j <= k$$

- Examine the ensemble averages and the cumulative average
- Also consider cumulative average when some of the initial batches are not considered
  - In other words, we start tabulating the data with batch m, with m >= 1

Steady State Analysis

- We can plot ensemble averages over time and see if there is a trend
- If necessary we can decide to eliminate some of the initial batches if they show too much bias

- Once we have done our runs we still need to analyze the data
  - Follow steps discussed earlier for transient systems