**Module-8:**

**Output Analysis of a Single Model**- Output analysis and types of simulation. Stochastic Nature of the Output Data. Measures of Performance and Estimation: Point Estimation and Confidence-Interval Estimation. Output Analysis for Terminating Simulations and Estimation of Probabilities. Output Analysis of Steady State Simulations: Initialization Bias, Error Estimation, Replications, Sample Size and Batch Means for Interval Estimation.

# Output Analysis for a Single Model

Its purpose is to predict the performance of a system or to compare the performance of two or more alternative system designs.⇐ Output analysis is the examination of data generated by a simulation. ⇐ This lecture deals with the analysis of a single system, while next lecture deals with the comparison of two or more systems. ⇐ Its purpose is to predict the performance of a system or to compare the performance of two or more alternative system designs. ⇐ Output analysis is the examination of data generated by a simulation. In brief

- Output analysis is the analysis of data generated by a simulation.
- How do we know if the result of a simulation is statistically significant?

**There are two types of simulations with respect to output analysis:**

**Terminating simulation:**

- Runs for some duration of time TETE, where E is a specified event that stops the simulation.
- Starts at time 0 under well-specified initial conditions.
- Ends at the stopping time TETE.
- Bank example: Opens at 8:30 am (time 0) with no customers present and 8 of the 11 teller working (initial conditions), and closes at 4:30 pm (Time TETE = 480 minutes).
- The simulation analyst chooses to consider it a terminating system because the object of interest is one day's operation.

A non terminating simulation is one that executes continuously.

**Non-terminating simulation:**

- Runs continuously, or at least over a very long period of time.

- Examples: assembly lines that shut down infrequently, telephone systems, hospital emergency rooms.
- Study the steady-state (long-run) properties of the system, properties that are not influenced by the initial conditions of the model.

**Stochastic Nature of Output Data**

**A stochastic simulation** is a simulation of a system that has variables that can change stochastically (randomly) with individual probabilities. [1]

Realizations of these random variables are generated and inserted into a model of the system. Outputs of the model are recorded, and then the process is repeated with a new set of random values. These steps are repeated until a sufficient amount of data is gathered. In the end, the distribution of the outputs shows the most probable estimates as well as a frame of expectations regarding what ranges of values the variables are more or less likely to fall in

Often random variables inserted into the model are created on a computer with a random number generator (RNG). The U(0,1) uniform distribution outputs of the random number generator are then transformed into random variables with probability distributions that are used in the system model.

- Simualtion result is generated for a given input data. Essentially a model is an input/output transformation.
- Since the input variables are random variables, the output variables are random variables as well, thus they are stochastic (probabilistic).
- Example 12.1 (Able and Baker, revisited)
  - Instead of a single run of simulation, four runs were conducted. The results are shown in Table 12.1 on page 431.
  - The utilization $\hat{\rho}_r$ and system time $\hat{w}_r$ were listed as 0.808,0.875,0.708, 0.842, and 3.74, 4.53, 3.84, 3.98.
  - There are two general questions we have to address by a statistical analysis of the observed utilization $\hat{\rho}_r$
    1. Estimation of the true utilization $\rho = E(\hat{\rho}_r)$ by a single value, called a point estimate.

2. Estimation of the error in our point estimate, either in the form of a standard error or confidence interval. This is called an interval estimate.

## Measures of Performance and Estimation

**Performance measurement** is the process of collecting, analyzing and/or reporting information regarding the performance of an individual, group, organization, system or component

The primary purpose of many simulation studies is the estimation of performance measures for the interesting system parameters through the most common statistic sample mean. To over the autocorrelation of observation data, batching observations is often taken in steady-state simulation analysis for estimating the variance of point estimators. But sample means as well as batch means, could not work well in the situation of much abnormal values or data with heavy-tail distributions. This paper gives the method of batch median for performance estimations in steady-state simulation. Comparisons between batch means and batch median are given both in theory and experimental studies. Empirical results illustrate that batch median loses very little in efficiency for normal correlated output, and retains its superiority in abnormal data over batch means.

Consider a set of output values for the same measure $Y_1, Y_2, \ldots, Y_n$ (e.g. delays of *n* different runs, or waiting times of *n* different runs). We want to have

- a point estimate to approximate the true value of $Y_i$, and
- an interval estimate to outline the range where the true value lies.

## Point Estimate vs. Interval Estimate

Statisticians use sample statistics to estimate population parameters. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

- **Point estimate**. A point estimate of a population parameter is a single value of a statistic. For example, the sample mean x is a point estimate of the population mean μ. Similarly, the sample proportion *p* is a point estimate of the population proportion *P*.
- **Interval estimate**. An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, *a* < x < *b* is an interval estimate of the population mean μ. It indicates that the population mean is greater than *a* but less than *b*.

### Point Estimates

A **point estimate** is a type of estimation that uses a single value, oftentimes **a sample statistic**, to **infer** information about the **population parameter**.

Let’s go through some of the major point estimates which include point estimates for the population **mean**, the population **variance** and the population **standard deviation**.

### Point Estimate for the Population Mean

So let’s say we’ve recently purchased 5,000 widgets to be consumed in our next manufacturing order, and we require that the average length of the widget of the 5,000 widgets is 2 inches.

Instead of measuring all 5,000 units, which would be extremely time consuming and costly, and in other cases possibly destructive, we can take a sample from that population and measure the average length of the sample.

As you know, the sample mean can be calculated by simply summing up the individual values and dividing by the number of samples measured.

$$\boldsymbol{Sample\ Mean}: \ \bar{X} = \frac{\sum x}{n}$$

### Example of Sample Mean Calculation

Calculate the sample mean value of the following 5 length measurements for our lot of widgets: **16.5, 17.2, 14.5, 15.3, 16.1**

$$\textit{\textbf{Sample Mean}}: \ \bar{X} = \frac{\sum x}{n} = \frac{16.5 + 17.2 + 14.5 + 15.3 + 16.1}{5} = 15.9$$

### Point Estimate for the Population Variance & Standard Deviation

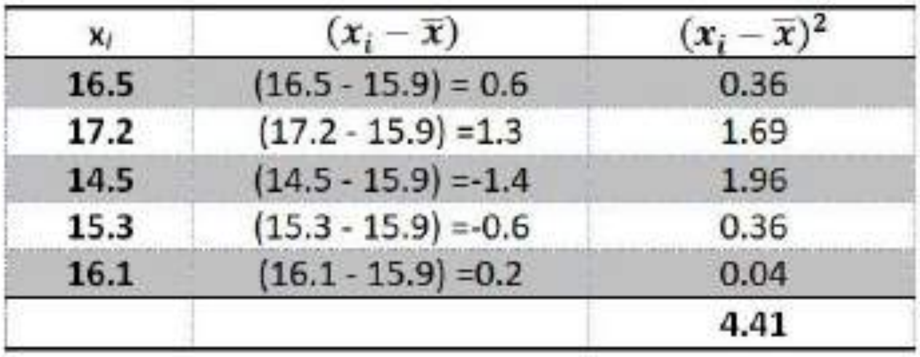
Similar to this example, you might want to estimate the variance or standard deviation associated with a population of product.

The point estimate of the population variance & standard deviation is simply the sample variance & sample standard deviation:

$$\textit{Sample Variance: } s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \qquad \& \qquad \textit{Sample Standard Deviation: } s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

### Example of Sample Standard Deviation

Let’s find the sample standard deviation for the same data set we used above: **16.5, 17.2, 14.5, 15.3, 16.1**



| $x_i$ | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ |
|---|---|---|
| **16.5** | (16.5 - 15.9) = 0.6 | 0.36 |
| **17.2** | (17.2 - 15.9) =1.3 | 1.69 |
| **14.5** | (14.5 - 15.9) =-1.4 | 1.96 |
| **15.3** | (15.3 - 15.9) =-0.6 | 0.36 |
| **16.1** | (16.1 - 15.9) =0.2 | 0.04 |
| | | **4.41** |

$$\textit{Sample Standard Deviation:}\ \ s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{4.41}{5-1}} = 1.05$$

## Confidence Intervals

Statisticians use a **confidence interval** to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.
- A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic* + *margin of error*.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the *sample statistic* + *margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

## Confidence Level

The probability part of a confidence interval is called a **confidence level**. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Here is how to interpret a confidence level. Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

## Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the **margin of error**.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

## Test Your Understanding

**Problem 1**

Which of the following statements is true.

I. When the margin of error is small, the confidence level is high.
II. When the margin of error is small, the confidence level is low.
III. A confidence interval is a type of point estimate.
IV. A population mean is an example of a point estimate.

(A) I only

(B) II only

(C) III only

(D) IV only.

(E) None of the above.

**Solution**

The correct answer is (E). The confidence level is not affected by the margin of error. When the margin of error is small, the confidence level can low or high or anything in between. A confidence interval is a type of interval estimate, not a type of point estimate. A *population* mean is not an example of a point estimate; a *sample* mean is an example of a point estimate.

## Output Analysis for Terminating Simulations

- Use independent replications, i.e. the simulation is repeated a total of *R* times, each run using a different random number stream and independently chosen initial conditions.
- Let $Y_{ri}$ be the *i*th observations within replication *r*, for $i = 1, 2, ... n_r$ and *r = 1, 2, ..., R*.
- For a fixed *r*, $Y_{r1}, Y_{r2}, ...$ is an autocorrelated sequence. For different replications $r \neq s$ $Y_{ri}$ and $Y_{sj}$ are statistically independent.
- Define a sample mean

$$\hat{\theta_r} = \frac{1}{n_r} \sum_{i=1}^{n_r} Y_{ri}, \quad r = 1, 2, ..., R$$

- There are *R* samples, so *R* sample means, the overall point estimate is

$$\hat{\theta} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta_r}$$

**Output Analysis for Terminating Simulations**

There are two types of simulations with respect to output analysis:

1. Terminating Simulation
2. Non-Terminating Simulation

A terminating simulation is one that runs for some duration of time TE where E is specified event that stops the simulation.

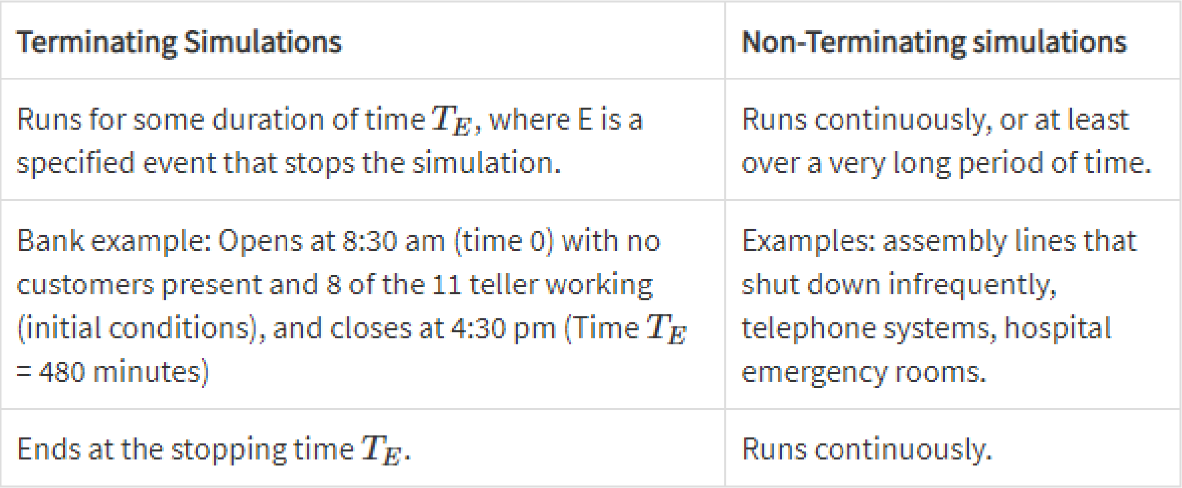
**Terminating simulation:**

- Runs for some duration of time TETE, where E is a specified event that stops the simulation.
- Starts at time 0 under well-specified initial conditions.
- Ends at the stopping time TETE.
- Bank example: Opens at 8:30 am (time 0) with no customers present and 8 of the 11 teller working (initial conditions), and closes at 4:30 pm (Time TETE = 480 minutes).
- The simulation analyst chooses to consider it a terminating system because the object of interest is one day's operation.

A non terminating simulation is one that executes continuously.

**Non-terminating simulation:**

- Runs continuously, or at least over a very long period of time.
- Examples: assembly lines that shut down infrequently, telephone systems, hospital emergency rooms.
- Study the steady-state (long-run) properties of the system, properties that are not influenced by the initial conditions of the model.



| Terminating Simulations | Non-Terminating simulations |
|---|---|
| Runs for some duration of time $T_E$, where E is a specified event that stops the simulation. | Runs continuously, or at least over a very long period of time. |
| Bank example: Opens at 8:30 am (time 0) with no customers present and 8 of the 11 teller working (initial conditions), and closes at 4:30 pm (Time $T_E$ = 480 minutes) | Examples: assembly lines that shut down infrequently, telephone systems, hospital emergency rooms. |
| Ends at the stopping time $T_E$. | Runs continuously. |

- Use independent replications, i.e. the simulation is repeated a total of *R* times, each run using a different random number stream and independently chosen initial conditions.
- Let $Y_{ri}$ be the *i*th observations within replication *r*, for $i = 1, 2, ... n_r$ and *r = 1, 2, ..., R*.
- For a fixed *r*, $Y_{r1}, Y_{r2}, ...$ is an autocorrelated sequence. For different replications $r \neq s$ $Y_{ri}$ and $Y_{sj}$ are statistically independent.
- Define a sample mean

$$\hat{\theta}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} Y_{ri}, \quad r = 1, 2, ..., R$$

- There are *R* samples, so *R* sample means, the overall point estimate is

$$\hat{\theta} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_r$$

**Estimation of Probabilities**

Simulation uses devices such as coins, number cubes, and cards to generate outcomes that represent real outcomes. Students may find it difficult to make the connection between device outcomes and the real outcomes of the experiment.

It defines the real outcomes of an experiment and specifying a device to simulate the outcome. Then, students are carefully led to define how an outcome of the device represents the real outcome, define a trial for the simulation, and identify what is meant by a trial resulting in a success or failure. Be sure that students see how a device that may have many outcomes can be used to simulate a situation that has only two outcomes, for example, how a number cube can be used to represent a boy birth (e.g., even outcome, prime number outcome, or any three of its digits).

Example 1 (10 minutes): Families

This first example begins with an equally likely model that simulates boy and girl births in a family of three children. Note that in human populations, the probabilities of a boy birth and of a girl birth are not actually equal, but they are treated as equal here. The example uses a coin in the simulation.

There are five steps in the simulation:

The first is to define the basic outcome of the real experiment (e.g., a birth).

♣ The second is to choose a device and define which possible outcomes of the device represent an outcome of

♣ the real experiment (e.g., toss of a coin, heads represents boy; roll of a number cube, prime number (P) represents boy; choice of a card, black card represents boy). The third is to define what is meant by a trial in the simulation that represents an outcome in the real

♣ experiment (e.g., three tosses of the coin represents three births; three rolls of a number cube represents three births; three cards chosen with replacement represents three births). The fourth is to define what is meant by a success in the performance of a trial (e.g., using a coin, HHT

♣ represents exactly two boys in a family of three children; using a number cube, NPP represents exactly two boys in a family of three children; using cards, BRB represents exactly two boys in a family of three children). Be sure that students realize that in using a coin, HHT, HTH, and THH all represent exactly two boys in a family of three children whereas HHH is the only way to represent three boys in a family of three children. The fifth step is to perform $nn$ trials (the more the better), count the number of successes in the $nn$ trials, and

♣ divide the number of successes by $nn$, which produces the estimate of the probability based on the simulation.

It may be useful to reiterate the five steps for every problem in this lesson so that students gain complete understanding of the simulation procedure.

Example 1: Families

How likely is it that a family with three children has all boys or all girls?

Let's assume that a child is equally likely to be a boy or a girl. Instead of observing the result of actual births, a toss of a fair coin could be used to simulate a birth. If the toss results in heads (H), then we could say a boy was born; if the toss results in tails (T), then we could say a girl was born. If the coin is fair (i.e., heads and tails are equally likely), then getting a boy or a girl is equally likely.

Pose the following questions to the class one at a time, and allow for multiple responses:

**How could a number cube be used to simulate getting a boy or a girl birth?**

♣ An even-number outcome represents boy, and an odd-number outcome represents girl; a primenumber outcome represents boy, and a non-prime outcome represents girl; or any three-number cube |digits represents boy, while the rest represents girl.

**How could a deck of cards be used to simulate getting a boy or a girl birth?**

♣ The most natural option is to allow black cards to represent one gender and red cards to represent the │other.

# Output Analysis for Steady-State Simulations

If initialization has in the point estimator has been reduced to a negligible level, the method of independent replications can be used to estimate point estimation variability and to construct a confidance interval.

If significant bias remains in the point estimator and a large number of replication are used to reduce point estimator variabilities the result confidence interval can be misleading.

The bias is not affected by the number of replications R, but by deleting more data (i.e increasing To) or extending the length of each run (i.e increasing TE)

Giving the run length, the number of replications should be as many as possible. Kelton is 1986 established that there is little valueto run more than 2S replications. So if time is available, make the simulation longer, instead of making more replication.

- Consider a single run of a simulation model whose purpose is to estimate a *steady state*, or *long run*, characteristics of the system.
- Assume $Y_1, Y_2, ...$ are observations, which in general are samples of an autocorrelated time series.
- The steady-state measure of performance $\theta$ to be estimated is defined by

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

- This result is independent of initial conditions, random number streams, ...

## Initialization bias in steady state simulations:

- Initial conditions may be artificial or unrealistic.
- Methods to reduce the point estimator bias include:

Intelligent Initialization:

- Initailize(start) the simulation in a state that is near expected of steady state( long run) conditions.
- Simulation takes some time to stabilize.

There are several methods of reducing the point estimator bias which is caused by using artificial and unrealistic initial conditions in a steady-state simulation.

1. Initialize the simulation in a state that is more representative of long-run conditions. E.g. use a set of real data as initial condition.
2. Divide the simulation into two phases, warm-up phase and steady state phase. Data collection doesn't start until the simulation passes the warm-up phase.

Consider the example

- *A set of 10 indep*endent runs, each run was divided into 15 intervals. The data were listed in Table 12.5 on page 453.
- Typicall we calculate average *within* a run. Since the data collected in each run is most likely autocorrelated, a different method is used to calculate the average *across* the runs.
- Such averages are known as *ensemble average*.

Several issues:

1. Ensemble average will reveal a smoother and more precise trend as the number of replications, *R*, is increased.
2. Ensemble average can be smoothered further by plotting a *moving average*. In a moving average each plotted point is actually the average of several adjacent ensemble averages.
3. Cumulative averages become less variable as more data are averaged. Thus, it is expected that the curve at left side (the starting of the simulation) of the plotting is less smooth than the right side.

4. Simulation data, especially from queueing models, usually exhibits positive autocorrelation. The more correlation present, the longer it takes for the average to approach steady state.
5. In most simulation studies the analyst is interested in several measures such as queue length, waiting time, utilization, etc. Different performance measures may approach stead state at different rates. Thus it is important to examine each performance measure individidually for initialization bias and use a deletion point that is adequate for all of them.

**Error Estimation**

The process of evaluating the uncertainty associated with a measurement result is often called **uncertainty analysis** or **error analysis**.

When we make a measurement, we generally assume that some exact or true value exists based on how we define what is being measured. While we may never know this true value exactly, we attempt to find this ideal quantity to the best of our ability with the time and resources available. As we make measurements by different methods, or even when making multiple measurements using the same method, we may obtain slightly different results. So how do we report our findings for our best estimate of this elusive **true value**? The most common way to show the range of values that we believe includes the true value is:

(measurement = (best estimate ± uncertainty) units

Let's take an example. Suppose you want to find the mass of a gold ring that you would like to sell to a friend. You do not want to jeopardize your friendship, so you want to get an accurate mass of the ring in order to charge a fair market price. You estimate the mass to be between 10 and 20 grams from how heavy it feels in your hand, but this is not a very precise estimate. After some searching, you find an electronic balance that gives a mass reading of 17.43 grams. While this measurement is much more **precise** than the original estimate, how do you know that it is **accurate**, and how confident are you that this measurement represents the true value of the ring's mass? Since the digital display of the balance is limited to 2 decimal places, you could report the mass as $m$ = 17.43 ± 0.01 g.

Suppose you use the same electronic balance and obtain several more readings: 17.46 g, 17.42 g, 17.44 g, so that the average mass appears to be in the range of 17.44 ± 0.02 g.

By now you may feel confident that you know the mass of this ring to the nearest hundredth of a gram, but how do you know that the true value definitely lies between 17.43 g and 17.45 g? Since you want to be honest, you decide to use another balance that gives a reading of 17.22 g. This value is clearly below the range of values found on the first balance, and under normal circumstances, you might not care, but you want to be fair to your friend. So what do you do now? The answer lies in knowing something about the accuracy of each instrument.

## TYPES OF ERRORS

Measurement errors may be classified as either **random** or **systematic**, depending on how the measurement was obtained (an instrument could cause a random error in one situation and a systematic error in another).

**Random errors** are statistical fluctuations (in either direction) in the measured data due to the precision limitations of the measurement device. Random errors can be evaluated through statistical analysis and can be reduced by averaging over a large number of observations (see standard error).

**Systematic errors** are reproducible inaccuracies that are consistently in the same direction. These errors are difficult to detect and cannot be analyzed statistically. If a systematic error is identified when calibrating against a standard, applying a correction or correction factor to compensate for the effect can reduce the bias. Unlike random errors, systematic errors cannot be detected or reduced by increasing the number of observations.

**Replication** is the repetition of an experimental condition so that the variability associated with the phenomenon can be estimated. ASTM, in standard E1847, defines replication as "the repetition of the set of all the treatment combinations to be compared in an experiment. Each of the repetitions is called a **replicate**."

Replication is not the same as repeated measurements of the same item: they are dealt with differently in statistical experimental design and data analysis.

For proper sampling, a process or batch of products should be in reasonable statistical control; inherent random variation is present but variation due to assignable (special) causes is not. Evaluation or testing of a single item does not allow for item-to-item variation and may not represent the batch or process. Replication is needed to account for this variation among items and treatments.

As an example, consider a continuous process which produces items. Batches of items are then processed or treated. Finally, tests or measurements are conducted. Several options might be available to obtain ten test values. Some possibilities are:

- One finished and treated item might be measured repeatedly to obtain ten test results. Only one item was measured so there is no replication. The repeated measurements help identify observational error.
- Ten finished and treated items might be taken from a batch and each measured once. This is not full replication because the ten samples are not random and not representative of the continuous nor batch processing.
- Five items are taken from the continuous process based on sound statistical sampling. These are processed in a batch and tested twice each. This includes replication of initial samples but does not allow for batch-to-batch variation in processing. The repeated tests on each provide some measure and control of testing error.
- Five items are taken from the continuous process based on sound statistical sampling. These are processed in five different batches and tested twice each. This plan includes proper replication of initial samples and also includes batch-to-batch variation. The repeated tests on each provide some measure and control of testing error.

Each option would call for different data analysis methods and yield different conclusions.

#### Sample Size and Batch Means for Interval Estimation

## Batch means for Interval Estimation in steady state simulation

One problem of replication method is that one has to delete some initial data. i.e. 'd' observations from some of the 'R' replications. ∴∴ We throw away a total number of 'dR' observations from all the observations.

- One disadvantage of replication is that data must be deleted on each replication.
- One disadvantageof a single-replication is its data tend to be autocorrelated.
- The method of *batch mean* divides the output data from one replication into a few large batches.
- Treat the means of these batches as if they were independentThe key isssue is that no commonly accepted method for choosing an accetable batch size *m*. This is actually one of the research areas in simulation.
  - Schmeiser found for a *fixed total sample size* there is little benefit from dividing it into more than *k* = 30 batches.
  - Although there is typically autocorrelation between batch means at all lags, the lag-1 autocorrelation $corr(\hat{Y_j}, \hat{Y_{j+1}})$ is usually studied to assess the dependence between batch means.
  - The lag-1 autocorrelation between batch means can be estimated using the method described earlier. They should not be estimated from a small number of batch means, i.e. we need to have large number of batches, though the size of batches could be small.
  - If the total sample size is to be chosen sequentially (i.e. choose one for one experiment, choose another one for improvement etc.), then it is helpful to allow the batch size and number of batches to grow as the run length increases.