Output Data Analysis (Part 4)

Dr.Çağatay ÜNDEĞER

Öğretim Görevlisi
Bilkent Üniversitesi Bilgisayar Mühendisliği Bölümü
&
...

e-mail : cagatay@undeger.com
cagatay@cs.bilkent.edu.tr

Output Data Analysis (Outline)

- Introduction
  - Types of Simulation With Respect to Output Analysis
  - Stochastic Process and Sample Path
  - Sampling and Systematic Errors
  - Mean, Standard Deviation and Confidence Interval
- Analysis of Finite-Horizon Simulations
  - Single Run
  - Independent Replications
  - Sequential Estimation
- Analysis of Steady-State Simulations
  - Removal of Initialization Bias (Warm-up Interval)
  - Replication-Deletion Approach
  - Batch-Means Method

Types of Simulation WRT Output Analysis

- Finite-Horizon Simulations
- Steady-State Simulations

Finite-Horizon Simulations

- Simulation starts in a specific initial state (e.g. empty, idle), and
- Runs until some termination event occurs (e.g. n jobs finished, working hours over).
- Life-time of process simulated is finite,
- So no steady-state behavior exists.
- Any parameter estimated from output depends on the initial state.

Finite-Horizon Simulations (Sample)

- Evaluation of a job process server:
  - Initial state:
    - Idle
  - Termination:
    - n jobs completed
  - Objective:
    - Estimate mean time to complete n jobs,
    - Estimate mean job waiting time.

Finite-Horizon Simulations (Sample)

- Evaluation of a military plan effectiveness:
  - Initial state:
    - Attact and defense are in their initial position, and operation is about to start.
  - Termination:
    - At most 25% of soldier left from either attact or defense forces.
  - Objective:
    - Estimate mean number of soldiers lost from attact and defense forces.

Steady-State Simulations

- The study of the long-term behavior of system of interest.
- A performance measure of the system is called a steady-state parameter.

Steady-State Simulations (Sample)

- Evaluation of a continuously operating communication system:
  - Objective:
    - Computation of the mean delay of a data packet.

Steady-State Simulations (Sample)

- Evaluation of a continuously operating military surveillance system:
  - Objective:
    - Computation of the mean ratio of threats that are not detected.

A Stochastic Process

- Counterpart to a deterministic process.
- Involves indeterminacy described by probability distributions.
- This means that;
  - Even if the initial condition is known,
    - There are many possibilities the process might go to, but some paths are more probable and others less.

A Stochastic Process

- Given a probability space , a stochastic process with state space X is a collection of X-valued random variables indexed by a set T (generally time).
- Often denoted as $\{X_t, t \in T\}$ or $\langle X_t \rangle$, $t \in T$.

A Sample Path

- A realisation of a stochastic process (one of the paths that can possibly occur).
- For instance, a sampled sequence of random variables, $X_1, X_2, X_3, \ldots, X_n$
- Each sample path has an associated probability to occur.
- In output data analysis,
  - State space X forms an output parameter
    - Whose sample paths are analyzed in order to reason about the process.

Sampling and Systematic Errors

- Every simulation experiment with random input generates random sample paths as output.
- Each path consists of a sequence of random observations.
- These sample paths include two kinds of errors that are:
  - Sampling error, and
  - Systematic error.

Sampling & Systematic Errors

- Sampling error:
  - The error caused by observing a sample instead of the whole population.
- Systematic error:
  - The error caused by biases (e.g. initial state of simulation) in measurement,
    - Which lead to measured values being consistently too high or too low, compared to the actual value of the measured parameter.

The Mean

- Expected value of a random variable, which is also called the *population mean*.
- For a data set, the mean is the sum of all the observations divided by the number of observations.

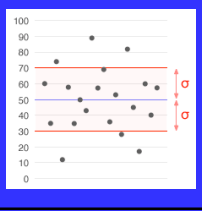
$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Standard Deviation

- A measure of the dispersion of a set of values sampled from a random variable.
- The mean is often given along with the standard deviation.
- The mean describes the central location of the data, and
- Standard deviation describes the spread.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \qquad s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

Entire population Sampled population

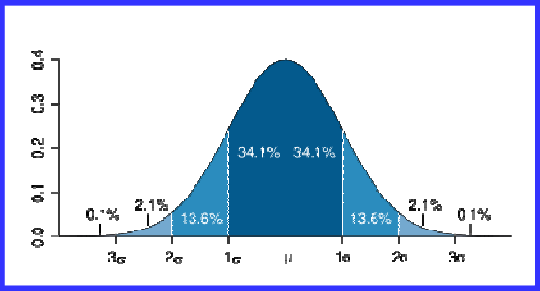


A data set with a mean of 50 and a standard deviation (σ) of 20

Standard Deviation

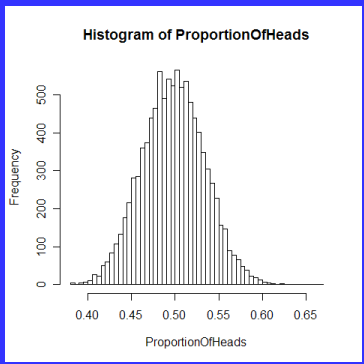
- In practice, it is often assumed that the data are from an approximately normally distributed population.
- This is ideally justified by the central limit theorem.

Dark blue is less than one standard deviation from the mean.



Central Limit Theorem

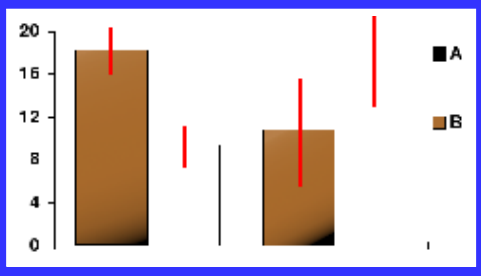
- Sum of a large number of independent and identically-distributed random variables will be approximately normally distributed.



Average proportion of heads in a fair coin toss, over a large number of sequences of coin tosses.

Confidence Interval

- A range of values centred on the sample mean $\bar{x}$ that is statistically known to contain the true mean μ with a given degree of confidence (usually taken as 95%).
- Used to indicate the reliability of an estimate.



• Top ends of the bars indicate observation means.

• The red line segments represent the confidence intervals surrounding them.

• The difference between the two populations on the left is significant.

Confidence Interval

- Specified by a pair (u,v),
  where $P(u \le \mu \le v) = 1-\alpha$
- $1-\alpha$ = confidence level or confidence coefficient
  where $0<\alpha<1$
- Confidence interval is computed by d
  where $P(\bar{x}-d \le \mu \le \bar{x}+d) = 1-\alpha$
- *So the interval for sample data is* $\bar{x} \pm d$

Confidence Interval (known σ)

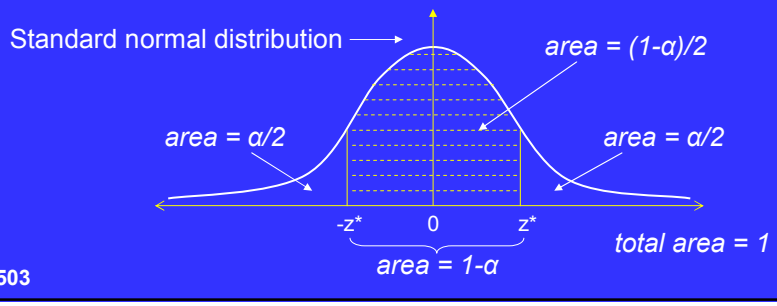
- The confidence interval for sample size n is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

σ ← true standard deviation

$z^* = 1.96$ for 95% confidence interval

$z^* = z_{1-\alpha/2}$ = point where area under right-half standard normal distribution is $(1-\alpha)/2$



Confidence Interval (unknown σ)

- In practice, true standard deviation for the population of interest is not known.
- Standard deviation is replaced by the estimated standard deviation S, known as standard error.
- x* (std.normal.dis) is replaced with t* (t-dis.).

$$\bar{x} \pm t^* \frac{S}{\sqrt{n}}$$

S ← estimated standard deviation

$t^* = t_{n-1,1-\alpha/2}$ = 1-α/2 probability value for t-distribution with n-1 degrees of freedom

A t-distribution Table

α = 0.2

α = 0.01

degrees of freedom (n-1)

df	$t_{0.1}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
2	1.89	2.92	4.3	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.6
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.	3.5
8	1.4	1.86	2.31	2.9	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.8	2.2	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.6	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.9
18	1.33	1.73	2.1	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.5	2.81
24	1.32	1.71	2.06	2.49	2.8
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.7	2.05	2.47	2.77
28	1.31	1.7	2.05	2.47	2.76
29	1.31	1.7	2.05	2.46	2.76
30	1.31	1.7	2.04	2.46	2.75
35	1.31	1.69	2.03	2.44	2.72
40	1.3	1.68	2.02	2.42	2.7
50	1.3	1.68	2.01	2.4	2.68
60	1.3	1.67	2.	2.39	2.66
70	1.29	1.67	1.99	2.38	2.65
80	1.29	1.66	1.99	2.37	2.64
90	1.29	1.66	1.99	2.37	2.63
100	1.29	1.66	1.98	2.36	2.63
200	1.29	1.65	1.97	2.35	2.6
300	1.28	1.65	1.97	2.34	2.59
400	1.28	1.65	1.97	2.34	2.59
∞	$z_{0.1}$	$z_{0.05}$	$z_{0.025}$	$z_{0.01}$	$z_{0.005}$
	1.28	1.645	1.96	2.33	2.58

Analysis of Finite-Horizon Simulations

- We would like to analyse the output of a simulation with the following properties:
  - Simulation starts in a specific initial state.
  - Runs until some termination event occurs.
  - Life-time of process simulated is finite.

Finite-Horizon Simulations (Single Run)

- Suppose that;
  - A simulation starts in a specific initial state,
  - Simulates a system until *n* output data $X_1, X_2, X_3, ..., X_n$ are collected.
  - Objective is to estimate $f(X_1, X_2, X_3, ..., X_n)$, where *f* is a "nice" function of data.
- For instance,
  - X_i may be transit time of unit *i* through a network, and
  - *f* may be average transit time for *n* jobs ($\overline{X}_n$).

Finite-Horizon Simulations (Independent Replications)

- *Unfortunately $\overline{X}_n$ is a biased estimator for μ and σ*
- *Since X_i's are usually dependent random variables making estimation of variance a difficult problem.*
- *To overcome the problem, multiple replications are required.*

- *Variance = Var(X) , σ^2_X or σ^2*
- *Estimated Variance = S^2_X or S^2*

Finite-Horizon Simulations (Independent Replications)

- *Assume that k independent replications of the system are run.*
- *Each replication starts with the same initial state.*
- *Each replication uses a different non overlaping portion of random number stream. To do that;*
  - *Start the 1st replication with a random seed,*
  - *Initialize the seed of next replication with the last random number produced by the previous replication*

    (doing nothing will already satisfy that rule).

Finite-Horizon Simulations (Mean and Variance)

- *Assume that replication i produces the output data $X_{i1}, X_{i2}, \ldots, X_{in}$ then*

Sample mean for i^{th} replication will be $$Y_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

Sample mean will be $$\overline{Y}_k = \frac{1}{k}\sum_{i=1}^{k} Y_i$$

Sample variance will be $$S^2_k(Y) = \frac{1}{k-1}\sum_{i=1}^{k} (Y_i - \overline{Y}_k)^2$$

Finite-Horizon Simulations (Confidence Interval)

- If *n* and *k* are sufficiently large, confidence interval for approximate *1-α* will be

$$\bar{Y}_k \pm t_{k-1,1-\alpha/2} \frac{S_k(Y)}{\sqrt{k}}$$

← sample standard deviation

Finite-Horizon Simulations (Sequential Estimation)

- *For fixed number of replications (k), we can not control the error in estimation of the mean.*
- *To limit the confidence interval for the mean within a tolerans ±d,*
  - *k could be determined incrementally.*
  - *Run one replication at a time and stop at the first k* satisfying*

$$t_{k-1,1-\alpha/2} \frac{S_k(Y)}{\sqrt{k}} \le \sqrt{\frac{k}{k-1} d^2 - \frac{t_{k-1,1-\alpha/2}}{k(k-1)}}$$

Simplification with little lost

$$t_{k-1,1-\alpha/2} \frac{S_k(Y)}{\sqrt{k}} \le d$$

Analysis of Steady-State Simulations

- We would like to analyse;
  - Long-term behavior of system of interest
  - By examining its steady-state parameters.

Steady-State Simulations (Removal of Initialization Bias)

- For analysing any steady-state parameter,
  - A simulation should first need to be converged to a steady-state.
- But since we start a simulation from an initial state (e.g. empty, idle),
  - Simulation will have a bias (warm-up interval),
  - We will need to wait some time until it is converged to the steady-state.
- Therefore, our first problem will be to detect the point where convergence occurs.

Steady-State Simulations (Removal of Initialization Bias)

- Most commonly used method for reducing the bias of $\overline{X}_n$ is:
  - To identify *m (1≤m≤n-1)*, which is the index of point where convergence is about to occur, and
  - Truncate the observations $X_1,...,X_m$.
- Then the estimator for $\overline{X}_n$ will be

$$\overline{X}_{n,m} = \frac{1}{n-m} \sum_{i=m+1}^{n} X_i$$

Steady-State Simulations (Graphical Method of Welch)

- One of most popular graphical methods is proposed by Welch (1981, 1983).
- Suppose there is *k* replications, and *n* observations for each replication.

Steady-State Simulations (Graphical Method of Welch)

- For the j^{th} observation, the estimated mean is

$$\overline{X}_j = \frac{1}{k} \sum_{i=1}^{k} X_{ij}$$

- Method plots *moving averages* $\overline{X}_j(w)$ of 1 to n observations on a graph for a given *time window w*.

$$\text{Moving average of } j^{th} \text{ obs.} = \overline{X}_j(w) = \begin{cases} \frac{1}{2w+1} \sum_{b=-w}^{w} X_{j+b} & w+1 \le j \le n-w \\ \frac{1}{2j-1} \sum_{b=-j+1}^{j-1} X_{j+b} & 1 \le j \le w \end{cases}$$

Steady-State Simulations (Graphical Method of Welch)

- For instance, when w = 2

$\overline{X}_1(2) = \overline{X}_1$

$\overline{X}_2(2) = 1/3\ (\overline{X}_1+\overline{X}_2+\overline{X}_3)$

$\overline{X}_3(2) = 1/5\ (\overline{X}_1+\overline{X}_2+\overline{X}_3+\overline{X}_4+\overline{X}_5)$

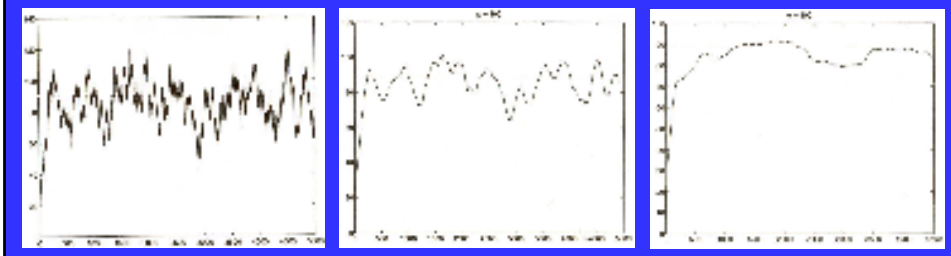
$\overline{X}_4(2) = 1/5\ (\overline{X}_2+\overline{X}_3+\overline{X}_4+\overline{X}_5+\overline{X}_6)$

...

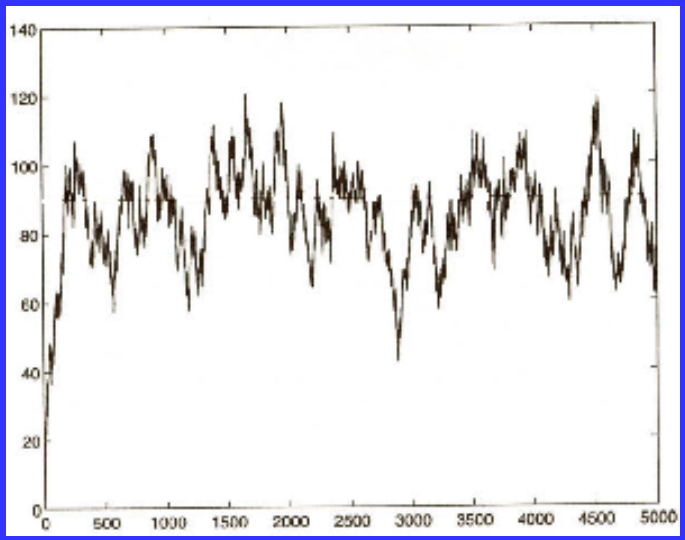
$\overline{X}_{n-2}(2) = 1/5\ (\overline{X}_{n-4}+\overline{X}_{n-3}+\overline{X}_{n-2}+\overline{X}_{n-1}+\overline{X}_n)$

Steady-State Simulations (Graphical Method of Welch)

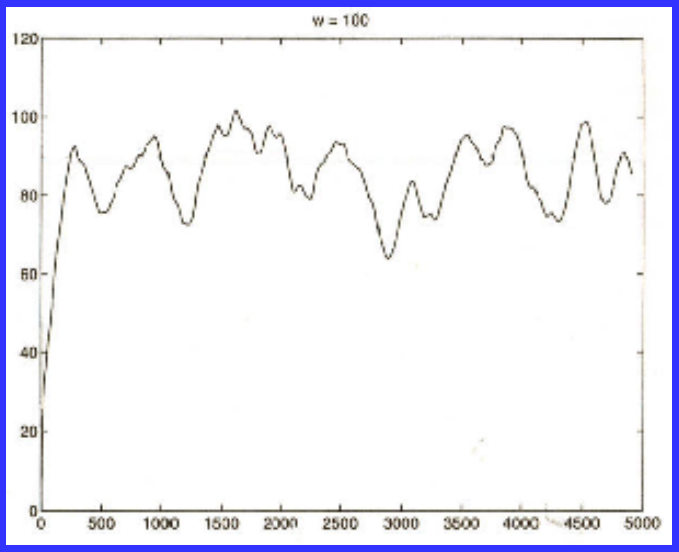
- If plot is reasonably smooth,
  - Cutoff *m* is chosen to be the value of *j* beyond which moving averages seems to be converged.
- Otherwise choose a different time window *w* and redraw the plot.



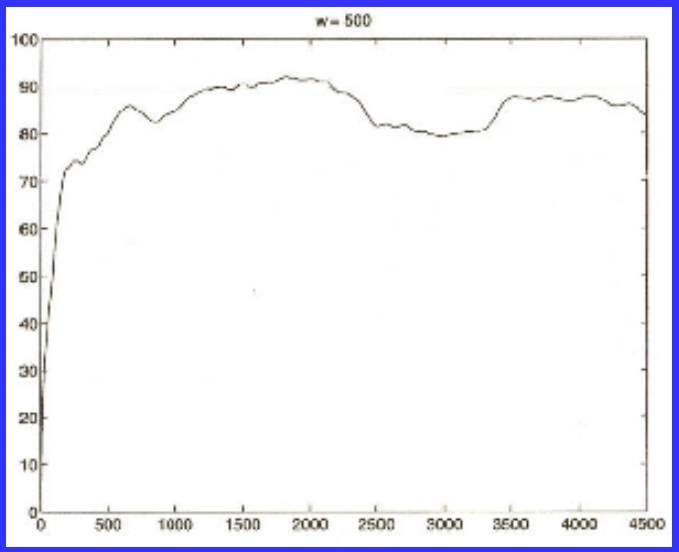
Steady-State Simulations (Graphical Method of Welch)



Steady-State Simulations (Graphical Method of Welch)



Steady-State Simulations (Graphical Method of Welch)



Steady-State Simulations (Replication-Deletion Approach)

- First determine initialization bias and cutoff *m* using any method such as Welch's.
- Run k independed replications each of length n observations, and
  - If possible, make use of runs from previous bias determination phase.
- Discard m observations from each replication.

Steady-State Simulations (Replication-Deletion Approach)

- Compute average of each replication

$$Y_i = \frac{1}{n-m} \sum_{j=m+1}^{n} X_{ij}$$

- Compute mean of replications

$$\bar{Y}_k = \frac{1}{k} \sum_{i=1}^{k} Y_i$$

- Compute confidence interval of replications

$$\bar{Y}_k \pm t_{k-1,1-\alpha/2} \frac{S_k(Y)}{\sqrt{k}}$$

Steady-State Simulations (Replication-Deletion Approach)

- Important characteristics:
  - As m increases for fixed n,
    - Systematic error due to initial conditions decreases.
    - But sampling error due to insufficient number of observations increases since variance is proportional to *1/(n-m)*.

Steady-State Simulations (Replication-Deletion Approach)

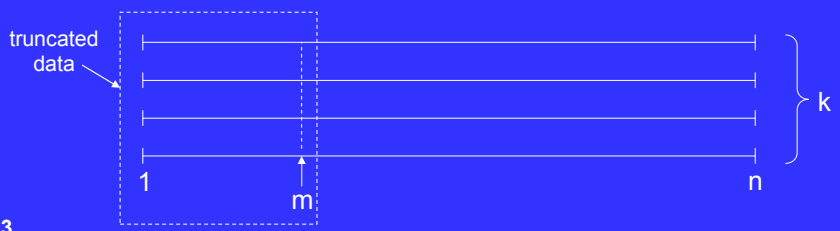
- Important characteristics:
  - As n increases for fixed m,
    - *Systematic error and sampling error decreases.*
    - *But runs take more time to finish.*

Steady-State Simulations (Replication-Deletion Approach)

- Important characteristics:
  - As k increases for fixed n and m,
    - *Systematic error does not change.*
    - *But sampling error decreases.*

Steady-State Simulations (Replication-Deletion Approach)

- Drawbacks:
  - Care must be taken to find a good cutoff m, and sufficiently large n and k.
  - Also there is potantially wasteful of data because of truncation from each replication.

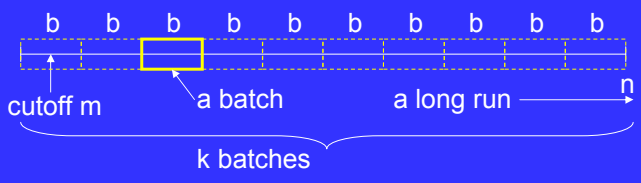


Steady-State Simulations (Batch-Means Method)

- One of the approaches that tries to overcome drawbacks of replication-deletion method.
- Owes its popularity to its simplicity and effectiveness.

Steady-State Simulations (Classical Batch-Means Method)

- Classical method:
  - Divides the output of a long simulation run with *n* observations into *k* number of batches with *b* number of observations in each batch (*b* = *n*/*k*),
  - Uses sample means of batches to produce point and interval estimators.



Steady-State Simulations (Classical Batch-Means Method)

- i^{th} batch consist of observations
  $X_{(i-1)b+1}$, $X_{(i-1)b+2}$, ... , X_{ib} for i = 1,2,...,k
- Mean of i^{th} batch is

$$Y_i(b) = \frac{1}{b} \sum_{j=1}^{b} X_{(i-1)b+j}$$

- Mean of entire run (grand batch mean) is

$$\bar{Y}_k = \frac{1}{k} \sum_{i=1}^{k} Y_i(b)$$

Steady-State Simulations (Classical Batch-Means Method)

- Variance of entire run is

$$S^2_k(Y) = \frac{1}{k-1} \sum_{i=1}^{k} (Y_i(b) - \bar{Y}_k)^2$$

- Confidence interval of entire run is

$$\bar{Y}_k \pm t_{k-1,1-\alpha/2} \frac{S_k(Y)}{\sqrt{k}}$$

← Standard deviation

Steady-State Simulations (Classical Batch-Means Method)

- Drawbacks:
  - Choice of batch size b is not easy.
  - If b is small,
    - Batch means can be highly correlated,
    - Resulting confidence interval will frequently have coverage below 1-α.
  - If b is large,
    - There will be very few batches, and
    - Potential problems with application of central limit theorem.

Steady-State Simulations (Classical Batch-Means Method)

- Selecting batch size & number:
  - Schmeiser (1982) stated that number of batches between 10 and 30 should suffice for most simulation experiments.
  - Chein (1989) showed that selecting *b* and *k* proportional to $\sqrt{n}$ performs fine in some conditions (SQRT Rule).
  - But in practice, SQRT rule tends to seriously underestimate variance for fixed n.

Steady-State Simulations (Overlapping Batch-Means)

- A variation of classical batch-means method.
- For a given batch size b, method uses all *n-b+1* overlapping batches.
- Therefore, i^{th} batch consist of observations $X_i , X_{i+1} , \ldots , X_{i-1+b}$ for i = 1,2,...,k
- Similar computations apply for mean and variance, but with different batch contents.