

Output analysis Overview

Properties of output variables

- Method of calculating of output
- Types of simulation output
- Point estimator
- The confidence interval

Terminating systems

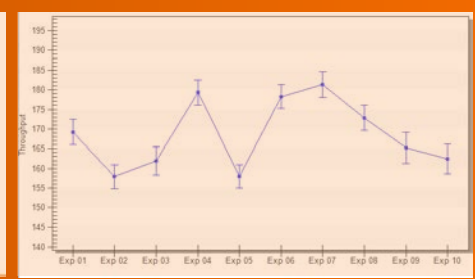
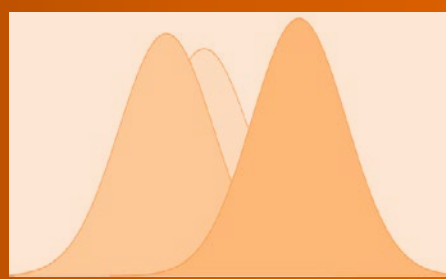
- Determining the number of observations
- Half-width of the confidence interval and the Two-phase method

Non-terminating systems

- Determining the truncation point
- Replication/Deletion approach
- Batch-means approach

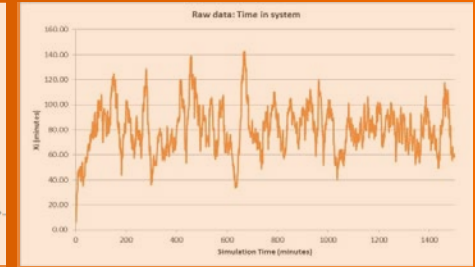
Evaluating and selecting scenarios

- Analysis of variance
- Applying ANOVA
- The Kim-Nelson procedure



Procedure KN

1. *Setup.* Select the overall desired $P(CS) = 1 - \alpha$, the value for δ , and common first stage sample size $n_0 \geq 10$. Set $\eta = \frac{1}{2} \left[\left(\frac{2\alpha}{k-1} \right)^{-2/(n_0-1)} - 1 \right]$.
2. *Initialization.* Let $I = \{1, 2, \dots, k\}$ be the set of scenarios still in contention, and let $k^2 = 2\eta(n_0 - 1)$. Obtain n_0 outputs X_{ij} ($j = 1, 2, \dots, n_0$) from each scenario i ($i = 1, 2, \dots, k$) and let $\bar{X}_i(n_0) = \sum_{j=1}^{n_0} X_{ij}/n_0$ denote the sample mean of the first n_0 outputs from scenario i .
For all $i \neq l$ calculate $S_{il}^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (X_{ij} - \bar{X}_i(n_0) - [\bar{X}_l(n_0) - \bar{X}_i(n_0)])^2$, the sample variance of the difference between scenarios i and l . Set $r = n_0$.
3. *Screening.* Set $I = I^{std}$. Let $I = \{i : i \in I^{std} \text{ and } \bar{X}_i(r) \geq \bar{X}_l(r) W_{\delta}(r), \forall l \in I^{std}, i \neq l\}$ where $W_{\delta}(r) = \max \left\{ 0, \frac{\delta}{2r} \left(\frac{k^2 S_{il}^2}{\delta^2} - r \right) \right\}$.
4. *Stopping rule.* If $|I| = 1$, then stop and select the scenario whose index is in I as the best. Otherwise, take one additional output $X_{i,r+1}$ from each system $i \in I$, set $r \leftarrow r + 1$, and go to Step Screening.



5. Output data analysis

R “Climate is what you expect; weather is what you observe.” (Morrison, 2019)

OUTPUT of stochastic models provide estimations of output parameters identified by the modeller. These are also referred to as *output variables*, *key performance indicators* (KPIs) or *performance measures*, and in the case of simulation optimisation, objective functions (see Table 2.2 and Chapter 6). In this chapter we study the statistical techniques used to configure simulation models to provide point and interval estimators for KPIs. We also study techniques to evaluate finite scenarios.

5.1 Simulation output analysis – overview

Systems containing stochastic elements/processes respond stochastically (see Kelton (1997)). A system is partially or completely described in terms of input and output (response) variables of a model. The following two concepts are important:

- *Response*: the consequence(s) of model logic and input data.
- *Statistical inference*: The methods by which one makes generalisations about populations.

When at least one input variable is stochastic, the model is not deterministic anymore, but has a random response that must be analysed with applicable statistical techniques. The analysis techniques applied in this chapter rely on aspects of statistics theory, of which a cornerstone is the Central Limit Theorem. It forms the basis of inference on expected values, and is stated next.

Theorem 5 (Central Limit Theorem). *If \bar{X} is the mean of a sample of size n taken from a distribution with mean μ and variance σ^2 , then $N(0, 1)$ is the limiting form of the distribution of*

$$Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \quad \text{as } n \rightarrow \infty. \quad (5.1)$$

The application of the theorem will become apparent later in the chapter.

A single response of a simulation model may not be used to draw conclusions from. It is similar to concluding that if person X has blue eyes, then every person in the world must have blue eyes. *A stochastic simulation executed on a computer is simply a computer-based statistical sampling experiment, and we rely on principles of Statistical Inference to analyse and draw conclusions from the results.*

Consider the simple deterministic system in Figure 5.1, where an entity arrives every minute, and is serviced for 30 seconds. It is obvious that the server will be occupied for 50% of the time, and that the time spent in the system is constant, that is 30 seconds. The time between exits will be 1 minute from the second entity and on.



Figure 5.1: Simple deterministic system

If the mechanisms in this system are changed to stochastic (arrivals and service) as shown in Figure 5.2, what will the effect on the exit times be?

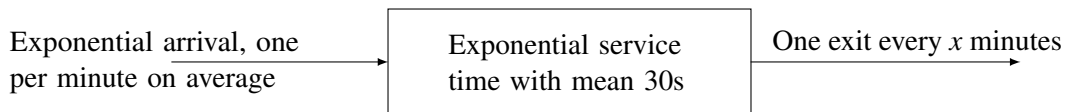


Figure 5.2: Simple stochastic system

It is not possible to predict the exit time with certainty anymore. This is what simulation is all about – to create a sufficient number of arrivals to draw a conclusion on the unknown parameter at the exit. The “sufficient number of arrivals” is determined with statistical analysis, as described later. A simulation model of a stochastic process yields different output values when compared to a deterministic system. The rule is often described as RIRO: Randomness in, randomness out. The principle of GIGO also applies in simulation, and more importantly, the principle of *RIRO* must be respected and correctly treated.

Figure 5.3 shows that we supply various forms of input to a stochastic simulation at different points in a model, and the resulting output parameters each has a response distribution.

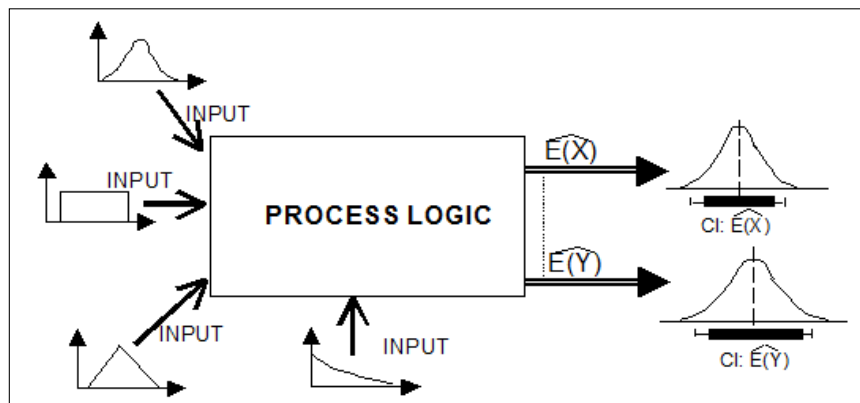


Figure 5.3: Input/output concept for a stochastic simulation model

Analysis of the output requires a different approach, which firstly depends on the type of system analysed (terminating *vs.* non-terminating), and other features like underlying statistical assumptions and the nature of the output variables, discussed next.

5.2 Properties of output variables

Output variables can be 1) classified according to the type of output they estimate, and 2) how they are observed.

5.2.1 Method of calculating of output

Given a set of observations $\{X_i\}$ generated with simulation, the methods of calculating estimated output are

- expected values ($\bar{X} = \sum_{i=1}^n X_i/n$),
- minimums ($\min\{X_i\}$),
- maximums ($\max\{X_i\}$),
- percentile (“How long do 95% of my customers wait at a specific point in the system?”), and
- proportions (“What proportion of set-up times take longer than 5 minutes?”).

5.2.2 Types of simulation output

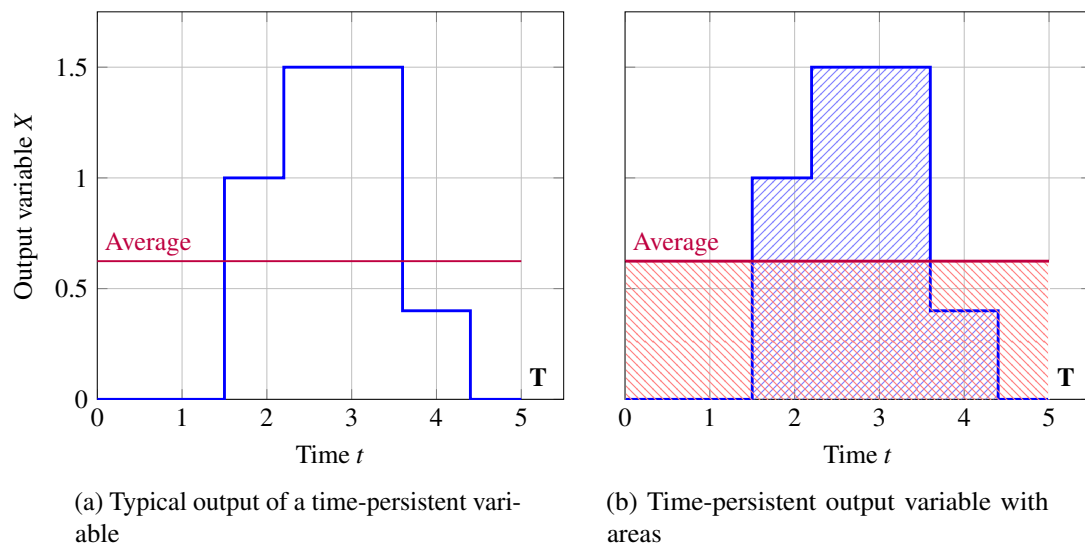
There are two main types of simulation output; these are:

1. **Observational** output: These simply contain single numbers which could be integer or real, for example
 - (a) the time it takes an operator to switch a jig,
 - (b) number of tasks completed,
 - (c) time to complete a task,
 - (d) time between arrivals.

In conjunction with the previous subsection, we determine the arithmetic mean of the observed values.

2. **Time-persistent** output: We often measure an output that persists at a level over time, for example the
 - (a) average queue length at a paypoint,
 - (b) utilisation of a resource,
 - (c) fluid level in a container,
 - (d) inventory level of a commodity.

An example of a time-persistent output variable X is shown in Figure 5.4a.



The level is observed over time from $t = 0$ until $T = 5$, then the average over that time length is calculated as the purple line. The average is the height of the horizontal line that defines a rectangle over the time of observation so that the red area and the blue areas in Figure 5.4b are equal. Note that the duration of the calculation starts at 0, even if the level is zero from $t = 0$ to $t = 1.5$.

An example of how to calculate the average for a time-persistent observation follows, in this case, the utilisation of two resources. In Figure 5.5 it is shown that two resources are operated over time, from $t = 0$ to $T = 15$. When no resource operates, the level is indicated as zero, while when any one resource is active, the active level is 1, and if both resources are operating, the level is 2. The levels on the graph change at discrete points in time, and the *combined utilisation* of the two resources is calculated as

$$\begin{aligned}
 \eta &= \sum \text{Subareas} / (\text{Number of resources} \times T) \\
 &= [((4-3) \times 1) + ((7-4) \times 2) + ((9-7) \times 1) + ((12-11) \times 1) + \\
 &\quad ((13-12) \times 2) + ((15-13) \times 1)] / (2 \times 15) \\
 &= 14/30 \\
 &= 0.467.
 \end{aligned}$$

Note that, as mentioned before, we divide in this example by 2×15 time units, because that is the duration of the time measurement for two resources – irrespective of the zero levels from $t = 0$ to $t = 3$. One can also think of the utilisation is the ratio of the area under the blue line divided by the rectangle enclosed by the vertical line $(0,0)$; $(0,2)$ and the horizontal line $(0,0)$; $(15,0)$.

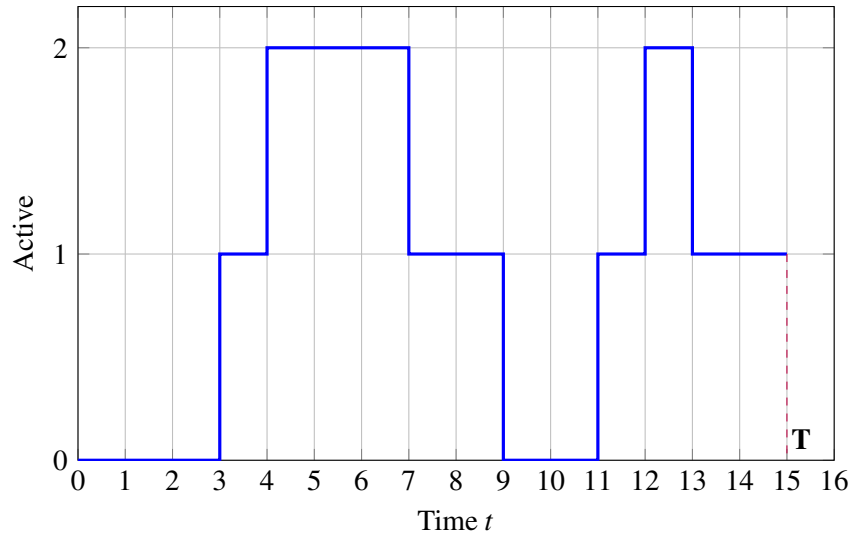


Figure 5.5: Time-persistent example for two resources that are operating or not

We focus on estimating *Expected values* of KPIs or output variables in this module, and typical variables are presented in Table 2.2 in Chapter 2. The objective of the output analysis is to estimate the value(s) of KPI(s) or output variable(s) of the system that is studied, and to describe them in terms of

1. *point* and
2. *interval estimators*.

These will be subsequently discussed.

5.2.3 Point estimator

The estimated mean values in Figure 5.9 are determined with simulation of a terminating system. The useful result from the Central Limit Theorem is that these means are approximately normally distributed (assuming that the number of observations per day is large), *i.e.* the random variable

\bar{X} is approximately normally distributed with mean $\bar{\bar{X}}$ and variance S^2 , where

$$\bar{\bar{X}} = \frac{\sum_{i=1}^n \bar{X}_i}{n} \quad (5.2)$$

and the sample variance $S_{\bar{X}}^2$ is an unbiased estimator of the population variance σ^2 , where

$$S_{\bar{X}}^2 = \frac{\sum_{i=1}^n (\bar{X}_i - \bar{\bar{X}})^2}{n-1} \quad (5.3)$$

with n = sample size ($n = 100$ in Figure 5.9). The estimated mean value calculated from (5.2) is a *point estimator* of the true population mean. The estimated variance of the mean follows from

$$S_{\bar{X}}^2 = \frac{S_{\bar{X}}^2}{n}. \quad (5.4)$$

R (5.4) can be proved as follows: We have a sample of \bar{X}_i , with sample size n . The observations are identically distributed (i.i.d.) and were sampled from the same distribution. Their variance is thus $\sigma_{\bar{X}}^2$. This is the *population variance*. Now, the sample variance is, under assumptions of i.i.d:

$$\begin{aligned} \text{Var}(\bar{\bar{X}}) &= \sigma_{\bar{\bar{X}}}^2 \\ &= \text{Var}\left(\sum_{i=1}^n \bar{X}_i / n\right) \\ &= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n \bar{X}_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(\bar{X}_i) \quad \text{due to independence} \\ &= \frac{1}{n^2} \times n \times \sigma_{\bar{X}}^2 \quad \text{because we summed } n \text{ times the same variance of the same distribution} \\ &= \frac{\sigma_{\bar{X}}^2}{n}. \end{aligned}$$

The sample variance decreases as n increases.

P Points to ponder: Suppose the observations are not normally distributed, but we do not know it. What will be the effect on the simulation result? If the observations are not independent, what will be the effect on the simulation result?

Next, an interval estimator for $\bar{\bar{X}}$ is explained.

5.2.4 The confidence interval

The result of (5.2) is called a *point estimator* of the population parameter μ . Another (more descriptive) estimator of a parameter is the *confidence interval*. This is an interval estimator of the population parameter, and *specifies a range in which the unknown population parameter is to be expected*. The confidence interval is also used to determine the number of replications required for the estimation of an output parameter, and we shall use the *t*-distribution to determine the confidence interval. The rationale behind the confidence interval and the use of the *t*-distribution is as follows:

- Recall from the CLT that $Z_n = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$, and $Z_n \xrightarrow{D} N(0, 1)$ as $n \rightarrow \infty$.
- It follows from the Central Limit Theorem that the random variable Z_n is approximately normal distributed if n is ‘large’.
- The mean $\bar{X}(n)$ is approximately normally distributed with unknown mean μ and variance σ^2/n , where σ^2 is the variance of the distribution of the population of X_i .
- In practice, σ^2 is usually unknown, but can be estimated with S^2 if n is large.
- The random variable $T_n = \frac{\bar{X}(n) - \mu}{\sqrt{S^2/n}}$ is approximately standard normal distributed ($N(0, 1)$).
- Therefore, $P(-z_{1-\alpha/2} \leq \frac{\bar{X}(n) - \mu}{\sqrt{S^2/n}} \leq z_{1-\alpha/2}) \approx 1 - \alpha$, where $z_{1-\alpha/2}$ is the upper $1 - \alpha/2$ critical point from the standard normal distribution.
- If n is ‘sufficiently large’, an approximate confidence interval for μ is

$$\bar{X} \pm z_{1-\alpha/2} \sqrt{S^2/n} \quad (5.5)$$

and the confidence interval half-width

$$h = z_{1-\alpha/2} \sqrt{S^2/n}.$$

- From the above, the question of what is ‘large’ arises.
- Firstly, if the X_i ’s are normally distributed, the random variable

$$T_n = \frac{\bar{X}(n) - \mu}{\sqrt{S^2/n}}$$

has a t -distribution with $n - 1$ degrees of freedom, and an *exact* $100 \times (1 - \alpha)$ confidence interval for μ is given by

$$\begin{aligned} CI &= \bar{X} \pm h \\ &= \bar{X} \pm t_{n-1, 1-\alpha/2} \sqrt{S^2/n}, \quad n \geq 2 \end{aligned} \quad (5.6)$$

where

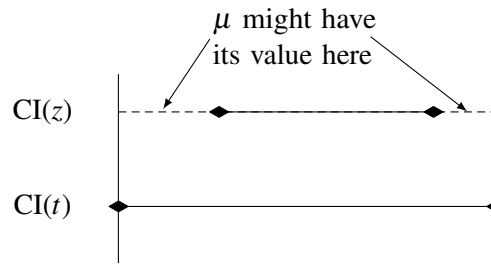
$t_{n-1, 1-\alpha/2}$ is the upper $1 - \alpha/2$ critical point from the Student t -distribution (upper critical values are shown in Table 6.4, Appendix A)

\bar{X} = Sample Mean

S^2 = Estimation of the population variance

n = Sample size, *i.e.* the number of observations.

- The X_i ’s are usually not exactly normally distributed (except where it can be assumed that they are approximately normally distributed due to the CLT, *e.g.* averages as output in a simulation study), so that the confidence interval from (5.6) is also approximate in terms of coverage.
- But $t_{n-1, 1-\alpha/2} > z_{1-\alpha/2}$, so that a wider confidence interval will result from (5.6) compared to that from (5.5), and the proportion of coverage will thus be closer to $1 - \alpha$ if the t -distribution is used. Refer to Figure 5.6.
- As $n \rightarrow \infty$, then $t_{n-1, 1-\alpha/2} \rightarrow Z_{1-\alpha/2}$, which will eventually differ very little. If n is small, the difference is significant, but at $n = 40$ the difference is only 3%.
- The interpretation of the confidence interval is as follows: If k confidence intervals are constructed for a given parameter, then it is expected that $(1 - \alpha) * k$ of these intervals

Figure 5.6: Coverage of the confidence interval: normal- and t -distribution

includes the true population parameter. This is known as the *coverage* of the confidence interval. Refer to Figure 5.7.

- We also conclude that the t -distribution must be used when the population variance must be estimated from the sample, *i.e.* when it is unknown.

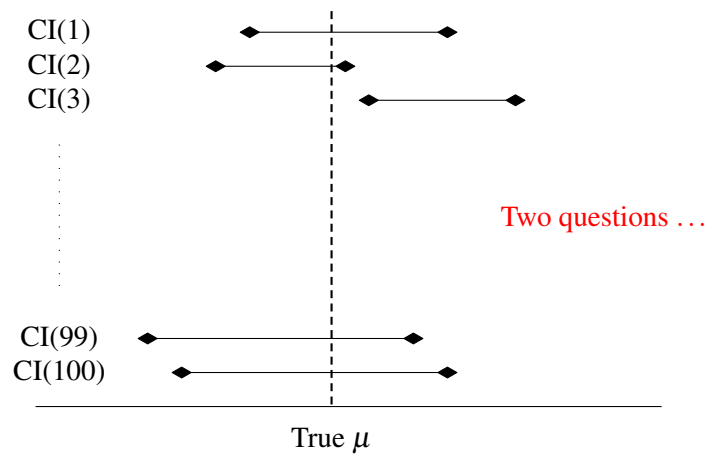


Figure 5.7: Interpretation of the coverage of a confidence interval

The coverage of the confidence interval and the error in estimating the true population mean is further explained in Figure 5.8. Note that ε indicates the error.

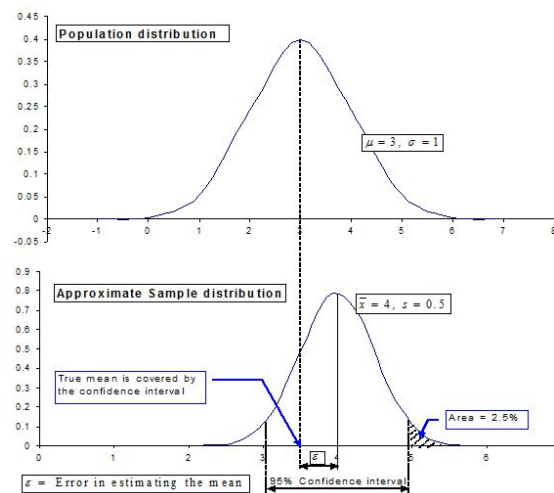


Figure 5.8: Error in estimating the true population mean

We shall now discuss the specifics of output analysis of terminating- and non-terminating systems when we simulate discrete-event, stochastic processes.

5.3 Terminating systems

A *terminating system* is defined as a system that starts in the empty state with operations idle, and after a logical event, ends in the empty state with operations idle again. The operations of a retail store stop after a specific time period and when there are no more customers to be served. The same principle applies to a restaurant. Every termination event usually supplies the analyst with an observation per output parameter, which may be assumed to be statistically independent of the observations from other replications during the same simulation run.

Suppose we are interested in the average time a party spends in a restaurant, and the true, but unknown value is denoted by μ . We can observe the time for every party per day since the restaurant is open until closing time. The arithmetic mean of these values can be calculated, which will result in one observation for that particular day. In simulation terminology, a model run that results in such an observation is called a *replication*. One replication in this case thus corresponds to a day in the restaurant. In terminating systems, each replication results in one observation, so that the ultimate objective is to determine how many replications are required to achieve a certain confidence level for a given output parameter.

The observation process can be executed for a number of days, say 100 days, and 100 averages will be available. The question is, did we make sufficient observations to draw a conclusion on this parameter, which is *Average time in restaurant*? These 100 values can be used to determine how many observations are actually needed to determine the parameter under study to a *specified level of confidence*. We therefore **estimate** the expected value of the parameter with a certain level of confidence, and this estimation process is a subset of *Statistical inference*. (See Gogg and Mott (1992), p. 9-3.) The process is described in Figure 5.9; the entry ‘Party x ’ denotes the time spent in the restaurant for the specific party. The value of μ is estimated in two ways: using a *point estimator* and an *interval estimator*.

Day 1	Day 2	Day 3	Day 4	...	Day 100
Party 1	Party 1	Party 1	Party 1	...	Party 1
Party 2	Party 2	Party 2	Party 2	...	Party 2
Party 3	Party 3	Party 3	Party 3	...	Party 3
\vdots	\vdots	\vdots	\vdots	...	\vdots
Party i	Party j	Party k	Party l	...	Party z

\bar{X}_1	\bar{X}_2	\bar{X}_3	\bar{X}_4	...	\bar{X}_{100}
-------------	-------------	-------------	-------------	-----	-----------------

Figure 5.9: Estimation of the mean of a parameter – terminating system

As stated earlier, simulation of stochastic systems is statistical sampling with the aid of a computer. A sufficient number of observations must therefore be made to draw a conclusion with a certain level of confidence on a parameter under study. This process is outlined next.

5.3.1 Determining the number of observations – terminating system

The main driver for the analysis is the distribution of the observations for the parameter under study. We will aim to limit the variance of the distribution to obtain a good estimation of the output parameter(s) under study. For this, we need the confidence interval half-width h and the *Two-phase method*, explained next.

5.3.2 Half-width of the confidence interval and the Two-phase method

We usually work towards a desired confidence interval width (a zero-width means that we have estimated the population parameter exactly), for which we require a number of simulation observations. One of several methods to determine the required number of observations (replications) is the *Two-phase method*, which is as follows (Pegden, R.E. Shannon, and R. Sadowski, 1995):

1. Phase 1: Do a trial run of 10 replications, *i.e.* $n = 10$. We therefore make 10 observations of the mean of each output parameter.
2. Estimate $\bar{\bar{X}} = \frac{\sum_{i=1}^n \bar{X}_i}{n}$ (using (5.2)) with the observations $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n$. Now estimate $S_{\bar{X}}^2 = \frac{\sum_{i=1}^n (\bar{X}_i - \bar{\bar{X}})^2}{n-1}$ from (5.3).
3. Determine the confidence interval half-width with

$$h = t_{n-1; 1-\alpha/2} \frac{S_{\bar{X}}}{\sqrt{n}}. \quad (5.7)$$

The value of t is the upper $1 - \alpha/2$ critical point on the cumulative Student t -distribution — make sure you understand what this means. It has for example a value of 2.262 at $n = 10$ and $\alpha = 0.05$. It can be calculated in MS-Excel 2010 with the formula “=T.INV(α , degrees of freedom)”, *i.e.* “=T.INV(0.975, 9)” will give 2.262158887, which is the upper critical t -value which has 97.5% of the area of the distribution with nine degrees of freedom to its left. This is for $\alpha=5\%$, so you must halve its value to get the two-tailed t -value. If one changes (5.7) to


$$h = t_{n-1; 1-\alpha} \frac{S_{\bar{X}}}{\sqrt{n}} \quad (5.8)$$

then one can use ‘=T.INV.2T(α , degrees of freedom)’, *i.e.* ‘=T.INV.2T(5%, 9)’ will also give $t=2.262158887$. Remember we use the ‘INV’ function because we need a t -value associated with a given probability.

4. If the calculated h is judged ‘too wide’ by the simulation analyst, they then select a smaller h called h^* , and the actual number of replications required to (hopefully) realise this narrower half-width is given by (Law and Kelton, 2000)

$$n^* = \left\lceil n \left(\frac{h}{h^*} \right)^2 \right\rceil \quad (5.9)$$

where h^* is the desired confidence interval half-width and n^* is the required number of replications. To use (5.9), the simulation analyst has to choose a target value for h^* . There are no recipes or rules for choosing a value for h^* , but it should at least be equal to or smaller than h (Why?). Students cringe when they have to choose a value without any guidance. If you understand the concept, it is easy. A smaller chosen value for h^* will result in a larger value for n^* which results in more replications and improved quality of estimation, but requires more simulation time.

 (5.9) can be proved as follows: let the variance estimator based on n observations be $S^2(n)$. Also, $n^* \geq n$ by definition, and we can assume that $S^2(n) \leq S^2(n^*)$, because

more observations make the variance smaller. Now,

$$\begin{aligned}
 h &= t_{n-1;1-\alpha/2} S(n) / \sqrt{n} \\
 h^* &= t_{n^*-1;1-\alpha/2} S(n^*) / \sqrt{n^*} \\
 \frac{h}{h^*} &= \frac{t_{n-1;1-\alpha/2} S(n) / \sqrt{n}}{t_{n^*-1;1-\alpha/2} S(n^*) / \sqrt{n^*}} \\
 &= \frac{t_{n-1;1-\alpha/2} S(n)}{t_{n^*-1;1-\alpha/2} S(n^*)} \cdot \frac{\sqrt{n^*}}{\sqrt{n}} \\
 h \times t_{n^*-1;1-\alpha/2} S(n^*) \times \sqrt{n} &= h^* \times t_{n-1;1-\alpha/2} S(n) \times \sqrt{n^*} \\
 h\sqrt{n} &\leq \sqrt{n^*} h^* \\
 \therefore n^* &\geq \left\lceil n \left(\frac{h}{h^*} \right)^2 \right\rceil.
 \end{aligned}$$

We introduce the “ \leq ” sign in the second last step because $t_{n-1;1-\alpha/2} \geq t_{n^*-1;1-\alpha/2}$ and $S(n) \geq S(n^*)$, so $t_{n-1;1-\alpha/2} S(n) \geq t_{n^*-1;1-\alpha/2} S(n^*)$. Because n^* is an integer, we round up as in (5.9), using the symbols \lceil and \rceil .

5. Phase 2: The simulation is run for n^* replications, and the calculations above are repeated (using (5.2), (5.3), (5.7) and (5.9)) to obtain a ‘final’ point and interval estimator, which are presented to the stakeholders of the study. The analysis is done *per output parameter*, and we take the maximum of n^* that is determined for each parameter, then run the simulation model for this number of replications.

General comments – number of replications

Why do we make 10 replications initially? Why not 15, 50, or 100? The number of 10 is a rule of thumb, but we also know that the t -distribution is useful for observations less than 30. We could make more than 30 observations, but the initial simulation runs may take too long. The use of 10 runs is thus an effort to compromise between simulation time (computer time) and the requirements of statistical procedures. If the observations are not normally distributed, then the results for the confidence interval hold for approximately 10 replications or more.

P Point to ponder: Does Statistics make uncertainty certain, or does it make us certain of the level of uncertainty?

Example: Calculating the desired number of replications

Assume the 10 observations given are means from 10 independent simulation runs:

\bar{X}_1	\bar{X}_2	\bar{X}_3	\bar{X}_4	\bar{X}_5	\bar{X}_6	\bar{X}_7	\bar{X}_8	\bar{X}_9	\bar{X}_{10}
16.818	8.124	18.895	1.879	4.394	11.654	1.086	2.353	16.500	5.435

From (5.2), the estimated mean is calculated as 8.714, and from (5.3) the estimated variance follows to be 45.968. Of course, $n = 10$, and the confidence interval half-width at $\alpha = 0.05$ is given by

$$\begin{aligned}
 h &= t_{n-1;1-\alpha/2} S_{\bar{X}} / \sqrt{n} \\
 &= 2.262 \times \sqrt{45.968/10} \\
 &= 4.850.
 \end{aligned}$$

Now select an $h^* < h$. The effect of the different values of h^* on n^* is shown in the following list.

h^*	2	1.5	1	0.75	0.25
n^*	59	105	236	941	3 764

A small h^* is desirable, but it can be seen that smaller values require many more replications.

P Point to ponder: Why $h^* < h$?

5.4 Non-terminating systems

Analysis of the simulation output of a non-terminating stochastic model requires more effort and knowledge than that of a terminating system. The main reasons are the presence of the transient phase, correlation, and the dependence of observations per parameter. The run length in terms of model time is also unknown. On the positive side, once certain data processing techniques have been applied, the remainder of the analysis is similar to terminating systems, as outlined in the previous section.

When a non-terminating model is simulated, observations are recorded as a sequential series of data for the duration of the simulation run. These observations are made as they occur, and they are not statistically independent due to the random number generation process and in-process variation. The transient phase also induces bias in the statistics. The techniques required to overcome these problems will now be discussed, and

- the moving average,
- correlation and the correlogram, and
- the replication/deletion and batching of observations

will be introduced.

5.4.1 Determining the truncation point

In the analysis of non-terminating systems it is usually desired to study the long-term or steady state behaviour of a system. (There are cases, however, where the transient phase is of particular interest: consider a proposed/new system – typical questions could be the length of the transient phase, the distribution of the parameters during that phase that is worst case, *etc.*). The truncation point is the point in simulation time where the steady state starts, and all data collected up to this point is discarded in analysis to eliminate bias.

The truncation point is determined by iteration and visual inspection using the moving average. Since we do not know where the truncation point is located, the simulation is run for an arbitrary length of simulated time. This might be too short, or sufficient. The output per parameter is used to draw a moving average. Suppose we study parameter X , and a number of observations were made over time, as follows:

Observation number	1	2	3	4	5
Time	0.2	3.3	5.1	5.8	7.1
Value of X_i	2.88	3.03	4.16	2.60	3.75
Average, $w=3$		3.357	3.263	3.503	

A moving average with window size $w = 3$ is used in this example. It means that the average of the first three values is taken (equal to 3.357, shown in column labelled '2'), then the average of values X_2, X_3 and X_4 is taken, and finally, the average of X_3, X_4 and X_5 is taken. A larger window size results in less averages. The effect the moving average has is to 'smooth' the original time-series signal (the values of X_i), so that a truncation point can be selected after visual

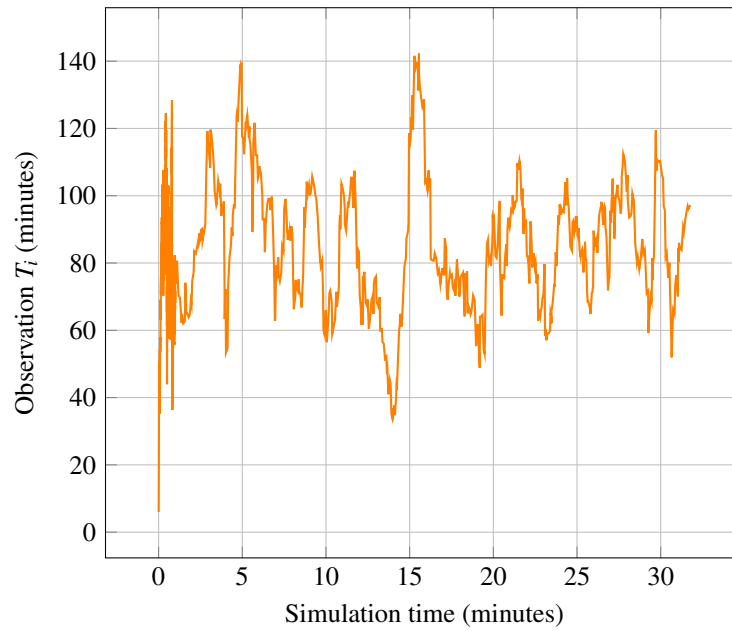


Figure 5.10: Typical output of a non-terminating system

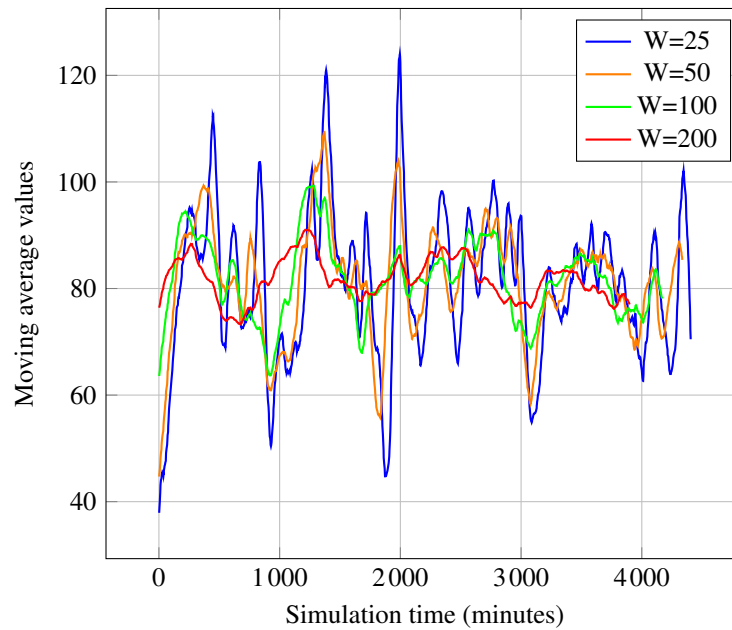


Figure 5.11: Moving averages with different window sizes

inspection. The observations shown in Figure 5.10 were smoothed using different window sizes, and the results are shown in Figure 5.11. The truncation point can be selected anywhere from 100 time units onwards, as it seems that the observations start to vary around a mean, while it seems as if the values increased from $t = 0$ up to 100. Choosing the truncation point is subjective and there are many correct answers for a given output series.

Once the truncation point is determined, the production runs can be executed with a warm-up period of length equal to l . There are several methods for the analysis of the steady state observations (see for example Chen and Kelton (2000)), but only two, namely the *Replication/Deletion*

and *Batch means*, will be discussed.

5.4.2 Replication/Deletion approach

The analysis in this method is similar to that for terminating systems, but the observations of the warm-up period are discarded. The method seeks to obtain statistically independent observations to make the application of the methods of the terminating system analysis valid.

1. Make N replications of length M , where M is at least $2l$.
2. The averages across replications are given by

$$\bar{Y}_j = \frac{\sum_{i=l+1}^M Y_{ji}}{(M-l)} \quad \text{for } j = 1, 2, \dots, N.$$

The Y_j uses only the observations from the j th replication of the steady state, that is $Y_{j,l+1}, Y_{j,l+2}, \dots, Y_{j,M}$.

3. The \bar{Y}_j 's are IID variables with $\mu \approx E(Y_j)$. $\bar{Y}(N)$ is an approximate unbiased point estimator for μ , and an approximate $100(1 - \alpha)$ CI is for μ is given by

$$\bar{Y}(N) \pm t_{N-1, 1-\alpha/2} \sqrt{\frac{S^2(N)}{N}}$$

where $\bar{Y}(N)$ and $S^2(N)$ are calculated from equations 5.2 and 5.3, respectively.

5.4.3 Batch-means approach

This method also seeks to obtain statistically independent observations to make the application of terminating system analysis possible. Only one replication is required with this method, as opposed to the Replication/Deletion approach; the simulation run therefore passes only once through the warm-up period.

Assume that the l observations of the warm-up period have been discarded, so that we deal only with observations X_{l+1}, X_{l+2}, \dots . Suppose the run produced m observations, then the observations are divided into n batches of length k . Batch 1 therefore consists of observations X_1, X_2, \dots, X_k , batch 2 consists of observations $X_{k+1}, X_{k+2}, \dots, X_{2k}$ and so forth. These observations are now averaged per batch, so that the n batches produce n averages $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n$, which form the new set of observations (also see Fishman (1978)). The overall sample mean is

$$\bar{\bar{X}}(n, k) = \frac{\sum_{j=1}^n \bar{X}_j(k)}{n} \quad (5.10)$$

which will be used as the point estimator for μ . The terms *warm-up*, *batches* and *steady state* are explained in Figure 5.12.

The main problem with this method is the choice of the batch size k . Many textbooks state that this method is acceptable if the batch size is *large enough*. But what is *large enough*? The correlation structure of the observations and the correlogram may provide a solution to this restriction, but before this concept is explained, additional background will be provided on *covariance* and *correlation*.

Covariance

The covariance between two random variables X and Y is a measure of their linear dependence, and is defined as

$$\sigma_{XY} = \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]. \quad (5.11)$$

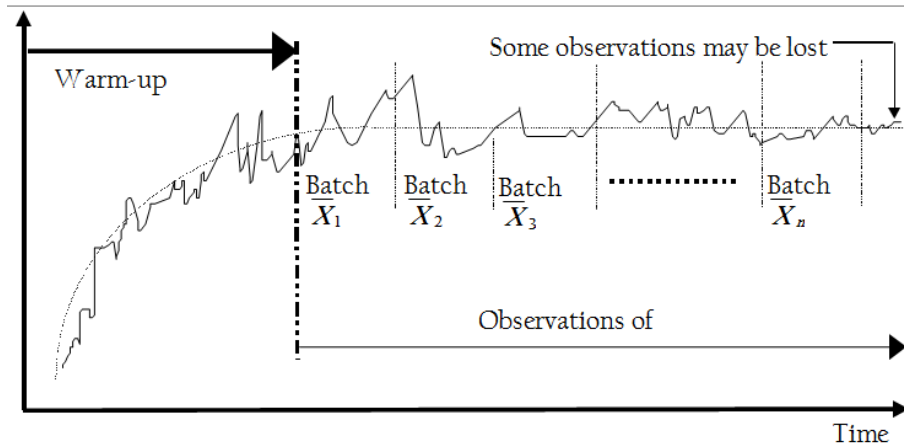


Figure 5.12: Discarding the transient phase and batching observations

- X and Y are uncorrelated if $C_{XY} = 0$, and if X and Y are independent, then $C_{XY} = 0$. The converse however, is *not generally true*.
- If $C_{XY} > 0$, then X and Y are positively correlated, so that $X > \mu_X$ and $Y > \mu_Y$ tend to occur together, and $X < \mu_X$ and $Y < \mu_Y$ occur together. Thus, if the one random variable is large, the other tends to be large (think of waiting times in a queue – if the customer in front of you waits a long time to be served, it is likely that the time you spend in the queue will also be long).
- If $C_{XY} < 0$, then X and Y are *negatively correlated*, so that $X > \mu_X$ and $Y < \mu_Y$ tend to occur together, and $X < \mu_X$ and $Y > \mu_Y$ occur together. Thus, if the one random variable tends to be large, the other tends to be small.

Correlation

The covariance σ_{XY} as a measure of dependence between two random variables is not dimensionless, which makes its interpretation difficult. If the random variables are measured in dimension x for example, then the covariance is measured in dimension x^2 . A dimensionless measure of linear independence is the *correlation* or *coefficient of correlation*, defined as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \quad (5.12)$$

and $-1 \leq \rho \leq 1$. Things to note:

- ρ_{XY} will have the same sign as σ_{XY} .
- If ρ_{XY} is close to +1, then the random variables X and Y are highly positively correlated.
- If ρ_{XY} is close to -1, then the random variables X and Y are highly negatively correlated.

These definitions imply that we have two sets of data from which we want to determine a covariance or correlation, for example X = mass of a person and Y = height of a person. If we sample a group of people by determining the mass of each, and measure the height of each, we could determine whether there is a correlation between a person's mass and height or not.

P **Point to ponder:** What is the difference between 'causation' and 'correlation'?

Covariance and correlation in a single data set

With simulation, we usually have a single set of observations per parameter under study, for example the flow time of patients through an emergency room. (We can of course have more

than one set of observations, which we want to study). It is often desired to determine if there is correlation among the observations in order to apply certain statistical operations, which assume independent observations. How would we determine this? The answer is to use the correlation coefficient as defined above with the single data set implicitly divided into subsets. This coefficient gives a measure of correlation/dependence between two observations and is used instead of the covariance, because it is dimensionless and therefore easier to interpret.

A different way to determine the covariance and correlation of a *single data set* must now be followed. The covariance in a single data set is defined as

$$C_j = \sum_{i=1}^{n-j} \frac{(X_i - \bar{X})(X_{i+j} - \bar{X})}{n-j} \quad (5.13)$$

where n is the number of observations. The sample correlation ρ_j is defined as

$$\rho_j = \frac{C_j}{\sigma^2} \quad (5.14)$$

where j is the so-called *lag*. The application of (5.13) and (5.14) is illustrated in the following paragraph.

Determining the correlation in a single data set

As mentioned earlier, we usually have a single set of observations per parameter under study which we want to analyse statistically, but we first have to know whether the observations are independent or not, or alternatively the degree of dependence on a scale of -1 to $+1$.

We now do not have two random variables X and Y that are each represented by a set of data, and we cannot use the strict definitions as defined above. We will have to work with data sets within our single data set (which reduces this explanation to one of explaining the use of indices of variables in a vector).

Suppose we have $n = 20$ observations, say random numbers between 1 and 10 inclusive. To determine the degree of dependence among the observations, we calculate the correlation. The process is as follows:

Calculate the sample mean with

$$\bar{X} = \sum_{i=1}^n \frac{X_i}{n}.$$

Estimate the *population* variance

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}$$

and calculate the covariance among the observations within the vector, applying the indices as indicated in

$$C_j = \sum_{i=1}^{n-j} \frac{(X_i - \bar{X})(X_{i+j} - \bar{X})}{n-j}. \quad (5.15)$$

This process in (5.15) is repeated for various lag sizes, and a great number of calculations are to be done to generate a correlogram when the number of lags is fairly large. It is often sufficient to draw the correlogram for 50 lags, but the final number depends on the problem. An example with a small sample in a spreadsheet with $j = 5$ correlations calculated is shown in Table 5.1; note the many numbers that are needed to be stored to obtain five results. It can be seen that, if there are say n observations in the data set under study, then for $j = 1$ one needs to make $n - 1$ subtractions of \bar{X} and $n - 1$ multiplications, then sum the terms. In the example given, this is executed 85 times.

Table 5.1: Example: Correlation calculations on a single data set

Obs. No.	Obs. i	$(X_i - \bar{X})(X_{i+j} - \bar{X})$					
		$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	
	1	7.984	0.132	-7.281	2.946	3.810	1.867
	2	6.830	-0.607	0.246	0.318	0.156	0.278
	3	0.942	-13.531	-17.496	-8.573	-15.301	17.512
	4	9.065	7.080	3.469	6.192	-7.087	-5.533
	5	9.750	4.486	8.007	-9.164	-7.154	9.443
	6	8.207	3.923	-4.490	-3.506	4.627	-2.120
	7	9.371	-8.014	-6.256	8.258	-3.783	6.250
	8	3.696	7.161	-9.452	4.330	-7.153	-5.003
	9	4.360	-7.379	3.380	-5.584	-3.906	13.357
	10	9.845	-4.462	7.371	5.156	-17.631	1.038
	11	5.295	-3.377	-2.362	8.077	-0.475	-3.673
	12	9.087	3.902	-13.344	0.785	6.067	-9.061
	13	8.377	-9.333	0.549	4.244	-6.338	4.911
	14	1.075	-1.878	-14.512	21.673	-16.793	-1.938
	15	7.057	0.854	-1.276	0.988	0.114	-0.705
	16	9.293	-9.855	7.636	0.881	-5.446	
	17	2.888	-11.404	-1.316	8.133		
	18	9.697	1.020	-6.302			
	19	7.068	-0.727				
	20	4.605					
Average $\bar{X} =$	6.725	Sum/ $(n - j) =$	-2.211	-2.968	2.656	-4.768	1.775
Pop. Variance =	8.025	$\rho_j =$	-0.275	-0.370	0.331	-0.594	0.221
$n =$	20						

The Correlogram

The correlogram is a graphical representation of the correlations against the lag numbers and is presented in a bar chart format. It is used to determine where the correlation in the data set diminishes to a satisfactory level *i.e.* close to 0. A typical correlogram is shown in Figure 5.13.

Note how the correlation values decrease with increasing lag number.

The correlogram is used to visually determine the lag number where the correlation becomes insignificant, *i.e.* close to 0. The need for this lag number will become apparent in the following paragraph. Note that the lags may or may not cross the zero line. The single lag where a zero crossing occurs may not be taken as the number where no correlation exists; a range of lags must be close to zero and a number in this range should be considered for the desired lag number. The first number in such a range is usually taken, provided the correlations further on are lower than the selected value.

It is also important to note that the window size, the value of j in (5.13), influences the correlogram. A small value for the window will result in a partial correlogram that will not reveal the approximate zero correlation lag. The maximum value for the window is also limited – if n observations were made, and the window size is j , it follows from equation 5.13 that only $n - j$ lags can be computed.

Determining the production run length for the Batch-means approach

The first step in analyzing a stochastic non-terminating model, a pilot run, must be executed to determine the truncation point and the approximate zero correlation lag number. The replication

length must be determined by iteration, or a very long run may be used but with a computer time penalty. The iteration process may be further extended if the correlogram does not show acceptable zero correlation levels for all parameters under study.

Once the lag number where correlation is acceptably small is identified, it can be used to determine the batch length (in simulation time), and consequently the actual length of the single replication for the production run. Suppose we select $j = 150$ in Figure 5.13, which means that the batch size is 150. As a rule, the batch size is selected at least one order of magnitude larger, *i.e.* 10 times larger than required. In this example, the batch size is thus $150 \times 10 = 1\,500$ observations. Suppose the duration of the pilot run was 1 200 simulation hours during which 800 observations of the parameter under study were made. Assume it is given that 800 observations required 1 200 hours of simulated time, so that 1 500 observations will require

$$\begin{aligned} \text{Replication length} &= \frac{1\,500 \text{ obs.} \times 1\,200\text{h}}{800 \text{ obs}} \\ &= 2\,250 \text{ simulation hours.} \end{aligned}$$

The analysis is now similar to that of a terminating system – a single run of $10 \times 2\,250 \text{ h} = 22\,500$ simulated hours will be executed to produce 10 batch-means. The warm-up period must also be added to this time. Suppose the warm-up period was determined to be 1 000 simulated hours, then the actual run length will be $22\,500 + 1\,000 = 23\,500$ simulated hours. The 10 observations are determined by calculating the average of each batch, that is 10 averages of the 15 000 observations produced by the 22 500 simulated hours run. The first average (observation) follows from the average of the first 1 500 observations, and the second observation from the average of the second set of 1 500 observations *etc.* The observations made during the warm-up period must of course, first be discarded.

A confidence interval can now be constructed per parameter, and the required number of replications can be determined. The production run will then be done for n^* times the time length per batch length, which is 2 250 simulated hours in this case. The warm-up period must once again be added to the overall time. It is recommended that the truncation point and the zero

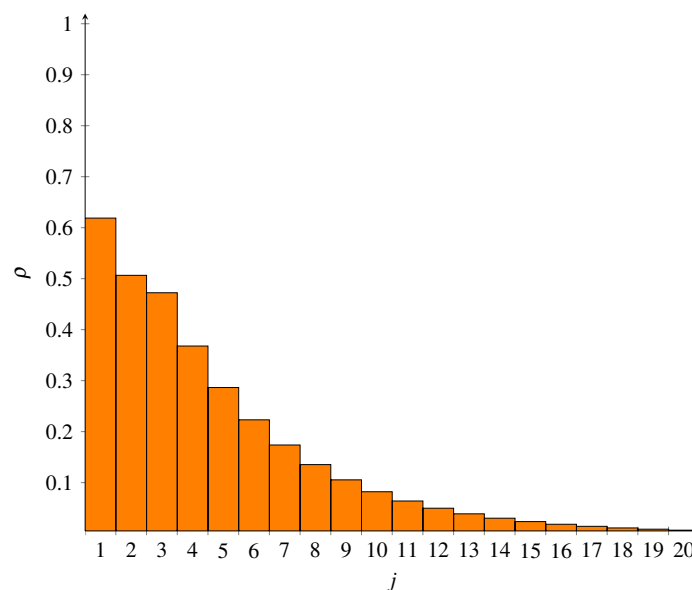


Figure 5.13: A typical correlogram

correlation lag be confirmed prior to the production run, by again analysing the output of the 10 batch-pilot run.

It is possible to test whether the batch size is sufficient or not. The von Neumann test as outlined in Banks (1998) can be used to determine if the batch means are independent.

Summary: Output analysis of non-terminating systems with Batch means

The analysis of a non-terminating system with the batch means approach is summarised in the following steps:

1. Do an initial run of arbitrary length.
2. Determine the length of the warm-up period per parameter with the method of moving averages. Iterate with the window sizes if necessary.
3. Do a long enough run so that sufficient data can be collected in the steady state. Use the data of the steady state to draw the correlogram.
4. If the length of the warm-up period as well as the approximate zero correlation lags can be determined, continue with the following steps, otherwise repeat the steps above with increasing run length.
5. Determine the number of observations required per batch by multiplying the approximate zero correlation lag number with 10. If more than one parameter is studied, the approximate zero correlation lag is the maximum of the lags of all parameters. If parameter 1 has an approximate zero correlation lag of 120, and parameter 2 has an approximate zero correlation lag of 95, then $120 \times 10 = 1\,200$ observations will be required per batch.
6. Determine the run time required to produce the required number of observations, and include the warm-up period.
7. Do a single replication pilot run of simulation time length as determined above.
8. Verify the truncation point and the approximate zero correlation lags.
9. Calculate the averages per batch.
10. Construct preliminary confidence intervals per parameter, and determine the required number of batches for each parameter's desired half-width.
11. Determine the production run length by multiplying the maximum number of replications required by the time required per batch, and add the warm-up period length.
12. Execute the single replication production run.
13. Calculate the averages of the observations per batch.
14. Draw confidence intervals per parameter.

P Points to ponder: What will the effect be on your mean estimation if you choose the warm-up length 'too short'? And 'too long'? Can it be 'too long'? What effect does the choice of the lag number j have on the estimation of the mean?

5.5 Evaluating and selecting scenarios

Simulation studies usually focus on a predefined set of scenarios to answer 'what-if' questions and one or more output parameters are used to determine which scenario is best. The number of scenarios could be small or large. Often, the simulation analyst formulates the scenarios, with or without stakeholders, then determines the best of these. A simple example is: do I need three, four, or five cashiers in a particular retail shop on Saturdays? This question implies three scenarios.

Defining scenarios thus means that we know what we want to investigate; in this case, it is usually a 'small' number. A 'large' number of scenarios usually constitutes simulation problems with such a large number of scenarios that it is impractical, or even impossible, to define them manually. For these problems, we need computerised methods to ease our task, which is the topic

of Chapter 6, but for now, the focus is on finding the best among a small number of scenarios. Several methods are available, collectively called *ranking and selection* (R&S) methods. For an overview of these methods, see Moonyoung Yoon (2017).

In this section, we shall study two of these methods, the first being *analysis of variance* (ANOVA), as it is the method Tecnomatix Plant Simulation uses. A second approach, namely the *Kim-Nelson method* (K-N method), will also be introduced and applied in this module using MS Excel.

5.5.1 Analysis of variance

Analysis of variance (ANOVA) is implemented in Tecnomatix Plant Simulation to assist the simulation analyst with finding the best scenario in a small set of scenarios. When analysing stochastic systems, the estimated parameter values may be numerically different, but they may not be *statistically significantly different*. A throughput value of $\bar{X}_1 = 22.4$ is not necessarily different from $\bar{X}_2 = 23.5$. ANOVA allows for finding scenarios that differ statistically significantly.

We use one-way ANOVA to determine the ‘better’ of a **small** number of scenarios. ‘One-way’ means a single output parameter is studied and compared. ‘Better’ is based on the estimated values of the output parameter, *e.g.* Throughput.

The general hypothesis test using ANOVA for k scenarios is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \neq \mu_k.$$

In our simulation studies, we test scenarios pair-wise: assume we have three scenarios to compare, then we compare Scenario 1 with 2, 1 with 3, and 2 with 3. For this purpose, the *paired t-test* is used while assuming unknown, unequal variances. The paired t-test is considered to be a special case of the general ANOVA.

R If three or more scenarios are to be compared, we should use the *F-test*, followed by a *post-hoc* test like Scheffe *post hoc* F test or the Tukey *post hoc* test to determine which pair of means is unequal. The paired t-test for more than two scenarios may result in compromising the Type I error rate. It could be used but the α -value should be corrected using the Bonferroni correction α/k , where k is the number of scenarios.

The critical test value is obtained by specifying a value for α , typically 5%. The test yields a p -value, and if $p < \alpha$, H_0 is rejected. This indicates that at least one μ_i is not equal to the others. The assumptions for the test are

- The samples are from normally distributed populations.
- Samples are independent.

The test is fairly robust and deviations from these assumptions do not affect results too much. The test values are determined as follows (Walpole and Myers, 1993):

H_0	Test statistic	H_1	Critical region
$\mu_1 - \mu_2 = d_0$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$	$\mu_1 - \mu_2 < d_0$	$t < -t_\alpha$
		$\mu_1 - \mu_2 > d_0$	$t > t_\alpha$
		$\mu_1 - \mu_2 \neq d_0$	$t < -t_{\alpha/2}$
			and $t > t_{\alpha/2}$
	$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$		
	$\sigma_1 \neq \sigma_2$ and unknown		

Tecnomatix does pair-wise comparisons to test pairs of means. If a test for any pair (μ_i, μ_j) , shows that $p < \alpha$, then $\mu_i \neq \mu_j$. If $\mu_i < \mu_j$ and we maximise, then Scenario j is better than

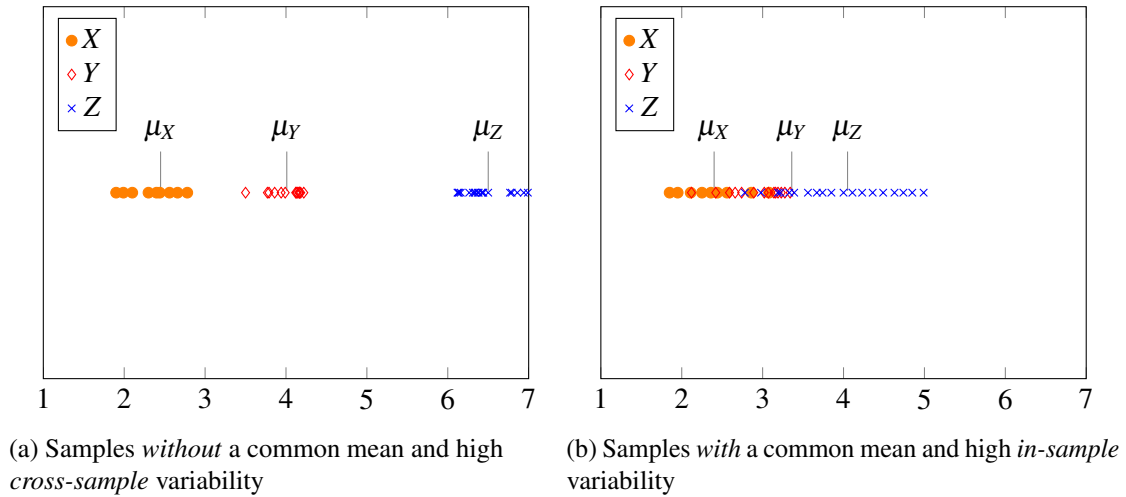


Figure 5.14: Cross-sample vs. in-sample variability

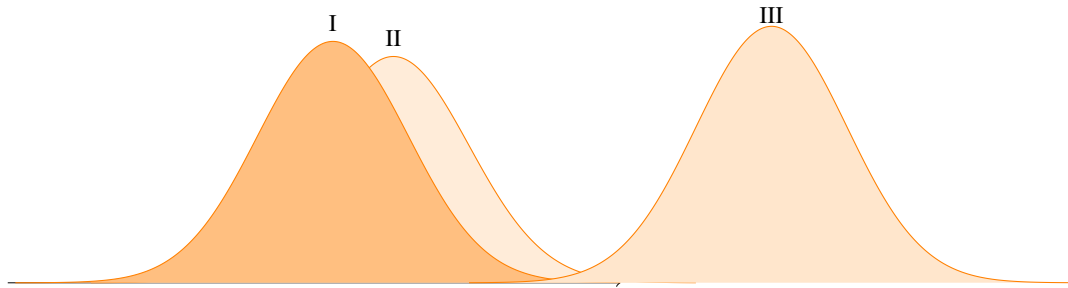


Figure 5.15: Samples with different distribution characteristics

Scenario *i*, and they differ statistically significantly.

The difference can be explained using Figure 5.14a. It shows that

- The populations do not have a *common mean*.
- The variability *between* samples is high.
- These imply that the populations *are different*, *i.e.* in simulation terms: the scenarios are different.

In Figure 5.14b, the populations **do** have a *common mean* while the variability *within* samples is high. These imply that the populations are not different, *i.e.* in simulation terms: the scenarios are *not different*.

Figures 5.14a and 5.14b show the difference between in-sample and cross-sample variation, and we conclude that if two scenarios are similar, *in-sample variation* is observed, while if two scenarios are different, their output exhibit *cross-sample variation*. The two figures can also be interpreted using Figure 5.15.

The two distributions labelled ‘I’ and ‘II’ are from the same population and exhibit in-sample variation, while the third distribution (labelled ‘III’) is from a different population, hence cross-sample variation between I, II *versus* III.

5.5.2 Applying ANOVA

Tecnomatix gives us a plot of the CI per experiment (scenario). It also gives us a pair-wise comparison of the estimated means of each experiment, and we need to determine which experiment is best. Consider an example with five experiments, where the real-life scenarios defined by them are increasingly more expensive, *i.e.* Exp1 is the cheapest to implement,

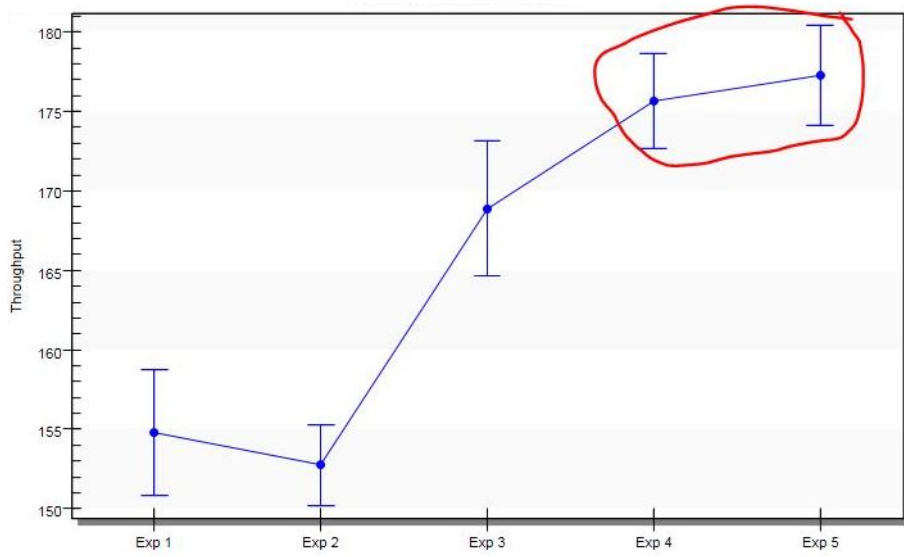


Figure 5.16: CI plot for ANOVA with highest scenario values

Table 5.2: Structure of p -value comparisons

	Exp 2	Exp 3	Exp 4	Exp 5
Exp 1	0.359	0	0	0
Exp 2		0	0	0
Exp 3			0.01	0.002
Exp 4				0.438

followed by Exp2, while Exp5 is the most expensive to implement. Suppose that, as an example, we simulate processes in a day-care hospital, and we want to maximise the *throughput* of the system, which is defined as *e.g.* Number of patients served per time period. A confidence interval plot for the five experiments is shown in Figure 5.16, with $\alpha = 5\%$. Each vertical bar shows a CI for an experiment, with each expected mean shown as a dot in the middle. The length of the line indicates the 95% confidence interval of the observations. Since we want to maximise throughput in this example, the values for Exp1 and Exp2 are too low for consideration and we reject these scenarios (experiments). The highest values are for Exp4 and Exp5, as shown in Figure 5.16.

It is clear that Exp5 has the highest numerical value for throughput, but is it really better than Exp4? We use the table with p -values from the ANOVA to determine that. The p -values associated with Figure 5.16 are shown in Table 5.2, for $\alpha = 5\%$. The direction of the ANOVA is symmetrical, *i.e.* comparing Exp4 and Exp5 is the same as comparing Exp5 and Exp4.

One can see that Exp1 and Exp2 have a common mean, as the p -value=0.359. Exp1 and Exp3 differ significantly, because the p -value<5%. This is also true for Exp(1, 4), Exp(1, 5), Exp(2, 3), Exp(2, 4), Exp(2, 5), Exp(3, 4) and Exp(3, 5). But Exp4 and Exp5 do not differ significantly because $p = 0.438 > \alpha$. We choose Exp4 as the best, since it has the highest throughput; although Exp5 has a numerically higher throughput, it is similar to Exp4 but more *expensive*. Remember the experiments were arranged to become progressively more expensive. We would thus not select Exp5, as it *yields the same output but is more expensive*. We can interpret this statement using an example: suppose the five scenarios represented appointment of one to five doctors in the day-care hospital, then Experiment 5 yields similar patient service, but with one more (expensive) doctor. The example shows the effect of in-sample variation (Figure

5.14b (Exp4 and Exp5), and cross-sample variation, *e.g.* Exp1 and Exp5 (Figure 5.14a).

Example of t-test

This example illustrates how the p -values in Table 5.2 were obtained. Suppose we have three scenarios, S1, S2 and S3, and the output parameter is *Cost*. Since we want to minimise cost, we seek the scenario which has the lowest cost and which differs statistically significantly from the others. The observations are shown in Table 5.3.

We need to do t-tests for (S1;S2), (S1;S3) and (S2;S3), and we test for no difference of observations, *i.e.* $\mu_1 - \mu_2 = d_0 = 0$. If the test fails, we can say that there is a significant difference between the two tested data sets. The results are shown in Table 5.4.

To calculate for example the t -statistic for (S1; S2), we use

$$\begin{aligned}
 t &= \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \\
 &= \frac{(26.3 - 27.0) - 0}{\sqrt{(0.357/10) + (1.875/10)}} \\
 &= -1.6212
 \end{aligned}$$

Table 5.3: Example observations for paired t-test

	S1	S2	S3
	27.0	28.3	24.2
	25.8	27.8	23.7
	26.6	27.8	23.6
	26.4	27.5	23.4
	24.9	24.2	21.4
	26.3	27.2	23.3
	26.8	27.5	23.5
	26.7	28.5	24.0
	26.1	25.4	22.7
	26.1	26.2	23.2
n_i	10	10	10
Mean	26.3	27.0	23.3
Variance	0.357	1.875	0.605

Table 5.4: Results for t-test example

	S1;S2	S1;S3	S2;S3
v	12	17	14
t-stat	-1.6212	9.5134	7.4622
p -value (two tails)	0.1309	0	0
t critical	2.1788	2.1098	2.1448

and the degrees of freedom with

$$\begin{aligned}
 v &= \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \\
 &= \frac{(0.357/10 + 1.875/10)^2}{\frac{(0.357/10)^2}{10 - 1} + \frac{(1.875/10)^2}{10 - 1}} \\
 &= 12 \text{ (rounded)}.
 \end{aligned}$$

Note that $n_1 = n_2 = n_3 = 10$, but it is not required, so the n_i may differ. The p -value can be found on statistical tables while taking the t -statistic value as absolute. If the table specifies a one-tailed value, then the obtained probability must be multiplied by 2. The t -test can also be done in MS-Excel: on the menu ribbon, click Data, Data Analysis, and choose t -Test: Two Sample Assuming Unequal Variances and follow the intuitive input requirements. In the next subsection, a procedure with a different approach to R&S is presented.

5.5.3 The Kim-Nelson procedure

The second ranking and selection procedure we study is the Kim-Nelson procedure (K-N procedure), due to Kim and Nelson (2001). There are two important concepts associated with the K-N procedure (and similar R&S procedures) which we must understand; they are:

- The *probability of correct selection*, $P(\text{CS})$. This is the probability of *correctly* selecting the scenario with the best output parameter value (*e.g.* lowest cost, highest throughput, shortest delay time). Since we are studying stochastic-driven scenarios, there is a probability of selecting the wrong one after analysis. This probability can be minimised but then many simulation replications must be done. We choose the probability of correct selection beforehand, and indicate it by $P(\text{CS}) = 1 - \alpha$, where typically $\alpha = 5\%$.
- The *indifference zone*, and indifference zone values, usually indicated by δ , δ_i or δ_{ij} , depending on the problem description. This value is specified by the simulation analyst beforehand, and is a value which indicates that the analyst is *indifferent* to two scenarios that have estimated means within the indifference zone value. The theoretical condition is $|\mu_1 - \mu_2| \leq \delta$, while the simulated condition would be $|\bar{X}_1 - \bar{X}_2| \leq \delta$. Of course, the value of δ must be carefully selected, keeping practical considerations in mind. Suppose we study the time a part spends waiting in a process, then it does not make sense to set $\delta = 0.01$ seconds, if the typical time is 3 hours and 20 minutes. So the simulation analyst must use judgement when choosing values for δ .

The K-N procedure is a *sequential* procedure, which means one estimates the output values using an initial number of replications n_0 , then sequentially does more replications per scenario, until sufficient observations are made to ensure $P(\text{CS})$. Estimates are thus calculated for n_0 replications, then for $n_0 + 1$ replications, and so on. The procedure is as follows (Kim and Nelson, 2001), assuming k scenarios:

Procedure KN

1. *Setup*. Select the overall desired $P(\text{CS}) = 1 - \alpha$, the value for δ , and common first stage sample size $n_0 \geq 10$. Set $\eta = \frac{1}{2} \left[\left(\frac{2\alpha}{k-1} \right)^{-2/(n_0-1)} - 1 \right]$.
2. *Initialisation*. Let $I = \{1, 2, \dots, k\}$ be the set of scenarios still in contention, and let $h^2 = 2\eta(n_0-1)$. Obtain n_0 outputs X_{ij} ($j = 1, 2, \dots, n_0$) from each scenario i ($i = 1, 2, \dots, k$) and let $\bar{X}_i(n_0) = \sum_{j=1}^{n_0} X_{ij}/n_0$ denote the sample mean of the first n_0 outputs from scenario i .

Table 5.5: Snapshot of observations X_{ij} for K-N procedure example

	i	$j \rightarrow$					
Scenario	1	7.20	5.61	8.44	6.30	10.48	...
Scenario	2	9.25	6.58	8.07	6.36	5.90	...
Scenario	3	8.32	7.34	16.81	9.89	10.96	...

For all $i \neq l$ calculate

$S_{il}^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (X_{ij} - X_{lj} - [\bar{X}_i(n_0) - \bar{X}_l(n_0)])^2$, the sample variance of the difference between scenarios i and l . Set $r = n_0$.

3. *Screening*. Set $I^{\text{old}} = I$. Let

$$I = \{i : i \in I^{\text{old}} \text{ and } \bar{X}_i(r) \geq \bar{X}_l(r) - W_{il}(r), \forall l \in I^{\text{old}}, l \neq i\}$$

$$\text{where } W_{il}(r) = \max \left\{ 0, \frac{\delta}{2r} \left(\frac{h^2 S_{il}^2}{\delta^2} - r \right) \right\}.$$

4. *Stopping rule*. If $|I| = 1$, then stop and select the scenario whose index is in I as the best. Otherwise, take one additional output $X_{i,r+1}$ from each system $i \in I$, set $r \leftarrow r + 1$, and go to Step *Screening*.

In Step 2 of the procedure, the subscript ‘ il ’ means ‘between Scenario i and Scenario l ’; the same applies to Step 3. The requirement in Step 2 “For all $i \neq l$ calculate” means we take all combinations of scenarios, for example, if we have three scenarios, then the combinations would be ‘1-2’, ‘1-3’ and ‘2-3’. The combinations are symmetric, so we do not consider ‘2-1’ and so on. We do not consider ‘1-1’, ‘2-2’ and ‘3-3’.

In Step 3, $W_{il}(r)$ determines how far the sample mean from Scenario i can deviate from the sample means of the other systems without being eliminated. If $W_{il}(r)$ becomes large, then Scenario i is eliminated. In the case of maximising the best scenario, the scenario that is finally selected has the largest true mean of all scenarios, and the true mean of the best scenario is also at least δ better than the second best.

Example: Kim-Nelson ranking and selection procedure

Three scenarios are present in this example. One thousand observations were created in advance from three normal distributions with means and standard deviations of $\mu_1 = 5, \sigma_1 = 1.8, \mu_2 = 10, \sigma_2 = 2$ and $\mu_3 = 12, \sigma_3 = 4$, respectively.

The 1 000 observations were created for the *purpose of this example* and may be too many – in practice, the simulation model will start with n_0 observations and sequentially make additional observations when necessary, which is the whole idea of the K-N procedure. A snapshot of the observations X_{ij} is shown in Table 5.5. We maximise, so we expect Scenario 3 to come out best because $\mu_3 = 12$.

The settings for Procedure K-N for the example are shown in Table 5.6. The value of η is calculated as in Step 1 of the procedure. Note that the indifference value was chosen as $\delta = 0.5$; this will/may dictate the outcome of the procedure: if it is chosen very small, many observations may be needed to distinguish scenarios, and it may become impossible for the procedure to do so. If it was chosen too large, the wrong scenario may be chosen as best because we ‘tell’ the procedure that we are very indifferent about the outcome. It is recommended to start with a relatively large value for δ , then decrease it as long as the optimisation does not take ‘too long’.

Table 5.6: Settings for the example of the K-N procedure

$\alpha = 5\%$	$k = 3$	$\delta = 0.5$	$n_0 = 10$	$\eta = 0.473$	$h^2 = 8.513$
----------------	---------	----------------	------------	----------------	---------------

The initial values for the K-N procedure are shown in Table 5.7.

Table 5.7: Initial values for K-N procedure example

	$\bar{X}_i(n_0)$	S_{il}^2
$il: 1,2:$	6.77	11.29
$il: 2,3:$	8.085	23.56
$il: 1,3:$	11.93	19.63

The progress of the K-N procedure is shown in Table 5.8. The scenario pairs were considered in the order (1,2), (2,3) and (1,3). The combinations are symmetric due to the squared term in Step *Initialisation*. First, the \bar{X}_s , $s = 1, \dots, k$ and the S_{il}^2 are determined, based on n_0 observations. These are shown in Table 5.7. Initially, 10 observations were made, then 11, then 12, until after 40 observations, only Scenario 3 was left in the set I . The note in the far right column indicates whether or not the scenario remains in set I , if not, it is not further considered.

The term $\bar{X}_i(r) \geq \bar{X}_l(r) - W_{il}(r)$ in Step *Screening* can be rearranged and written as $W_{il}(r) \geq \bar{X}_l(r) - \bar{X}_i(r)$. Scenario i stays as long as this is true, otherwise it is removed from the set I . In Table 5.8, at $r = 27$, $\bar{X}_1 = 5.92$, $\bar{X}_2 = 8.61$, and $W_{12}(27) = 2.57$. Because $W_{12}(27) = 2.57 \not\geq 8.61 - 5.92 = 2.69$, Scenario i is removed from I .

In the row at $r = 40$, the value of $W_{12} = 3.56$ is striked through, because it is not considered in the screening; only $W_{23} \geq \bar{X}_3 - \bar{X}_2$ is considered. In this row, $W_{23} = 2.93 \not\geq \bar{X}_3 - \bar{X}_2 = 3.71$, and Scenario 2 is eliminated, leaving Scenario 3 as the winner.



 Why bother with Procedure K-N if we have ANOVA?

Table 5.8: The sequential progress of the K-N procedure example











r	\bar{X}_i	$W_{il}(r)$	Stay/Out?
10	6.77	9.36	Stay
	8.08	19.80	Stay
	11.93	16.46	Stay
11	6.76	8.17	Stay
	8.21	17.32	Stay
	11.91	14.39	Stay
...
27	5.92	2.57	Out
	8.61	5.63	Stay
	12.43	4.65	Stay
...
40	5.92	2.57	Out
	8.91	3.56	Out
	12.62	2.93	Stay

There are many more procedures to rank scenarios and select the best, and the R&S field is still being actively researched.

 “Statistics means never having to say you’re certain” (Senn, 2018).

Survival kit:

If you want to successfully wade through the output analysis assessment jungle, make sure the following tools are in your bag:

-  1 Know the difference between a TS and NTS.
-  2 Know what is a point and interval estimator.
-  3 Interpret the confidence interval.
-  4 Know how to calculate n^* for a TS.
-  5 Know the two NTS analysis methods.
-  6 Find the truncation point for each output parameter of an NTS.
-  7 Know and use correlation when simulating an NTS and apply the correlogram.
-  8 Determine the batch size per output parameter of an NTS.
-  9 Determine n^* for an NTS.
-  10 Determine the best in a small set of scenarios using the t-test (manually) and the *properties* of the K-N procedure (MS-Excel).