9.2 Output analysis.

The output or result of a simulation run for a given data (initial values of variables and parameters) must be analyzed to make decisions or to compare it with results produced by different data, or by a modified model or with real data from the simulated system.
These results appear usually as time series of one or many variables, as numerical values representing statistical summaries, terminal or critical values of variables.
It may also appears as characters, pictures, sounds or written statements.
In **deterministic models** the results of a run for a given data are unique. Thus, the results of a run may be directly used for decisions and comparisons.
If the model has random variables, the same model **with the same data** may give different results in different runs, because different values of the seeds of the random numbers used in different runs produce different values of the random variables. In a simulation model processed by a computer, which is a deterministic machine, the result is, strictly speaking, deterministic. However, as we have seen in 6.2.2.c, a simulation model of a system with stochastic processes have two parts. One part **imitated the structure of the system** using certain variables and relations among them. Another part **imitated the stochastic characteristics of the system** giving values to these variables based in a **stream of numbers** that imitates a random sequence (pseudo random numbers). This stream is generated from a **seed** that is the first number of the stream (see 6.3.2). In a repetition or **replication** of a run, the data are the same of the previous run, but the seed of the random numbers is different. Therefore the output appears as a result of a random process: The sequence of results of a variable is different in each replication. We have seen that in a model several streams of random numbers may be used for different variables. All or some of them are changed in different replications of the run. (See Zeigler 1975)
The results are expressed as values of random variables and the analysis must be made by statistical decision and comparison methods.
In this section the special problem of **stochastic models** (models using random variables) will be treated. The processes described by these models are **stochastic processes.**

9.2.1 Some definitions about stochastic processes.

Before discussing output analysis it is convenient to remember some concepts of **stochastic processes**. In an stochastic process the values of the variables are **random values** given by a time series: $y_1, y_2, ..., y_m$ corresponding to times: $t_1, t_2, ..., t_m$. The last one is an increasing sequence of real values. Each value of y_i is obtained by a **random experiment**. In a random experiment the result or outcome may be different when repeated under the same conditions. Therefore, if we repeat the stochastic process the sequence of the values y_i is, in general, different. Note that the value of the variable may be a number, a logical value or a member of a definite set. If it is not a number, it may be convenient to use as result the value of a numerical function of the non numerical output. Thus, the outcome of a dice throw may be represented by an number of the points of the face. In the case of a coin the function may assign, for instance, 1 to head and 0 to tail.

Example 9.2.1. A dice is thrown 20 times. The number of points is a stochastic variable. For instance:
6,3,2,3,6,4,3,5,5,1,3,6,3,5,2,6,3,5,4,1.

A queue in an attending window is observed at 1,2,3,..,20 periods of 10 seconds, the observed values were:
0,1,3,3,4,4,4,5,6,6,4,4,3,3,2,2,2,4,5,6.

A stochastic process is called **stationary** if the **joint distribution of probability** of its variables remains the same in the time.

Example 9.2.2. An stochastic process produces pairs of results by the following experiment:
A token is extracted from a bag A which contain 12 tokens with the color red and 8 with blue.
If the result is red a token is extracted from a bag B with 6 tokens with the number 1, 30 with the number 2 and 24 with the number 3.
If the result is blue a token is obtained from a bag C with 28 tokens with the number 1, 40 with the number 2 and 12 with the number 4.
An outcome may be for instance: (red,1), (blue,3), (blue,2), (red,1),……

Example 9.2.3. In a simple teller in which, for example, the mean times between arrivals change with time, the values of the queue length have a variable probability distributions (see example 6.2.2.a). The process in not stationary.

The process of the example 9.2.2 is stationary. See Exercise 9.2.1.

In a stationary process the probability distribution of each variable is also constant.
But the reciprocal may be not true: **the distribution of each variable may be constant and the process may be non stationary.**

Example 9.2.4. In a process of throwing 2 separated coins there are two random variables, one for each coin. The possible outcomes may be represented by the values 0 and 1. The joint probability function is: $f(0,0)=1/4 \quad f(0,1)=1/4 \quad f(1,0)=1/4 \quad f(1,1)=1/4$. **(1)**
The probability function for the first coin is: $f_1(0)=1/2 \;\; f_1(1)=1/2$, and for the second one: $f_2(0)=1/2 \;\; f_2(1)=1/2$.
In a process in which the coins, before being thrown, are stuck each other by a point in the border with the sides opposed, in such a way that when one exhibits toss the other exhibits head, the probability functions f_1 and f_2 remain the same, but the joint probability will be different, because the joint probability when the coins are stuck are: $f(0,0)=0 \quad f(1,0)=1/2$
$f(0,1)=1/2 \quad f(1,1)=0$.
Consider a process in which the two coins are sometimes free and sometimes stuck, for example:
Stuck, free, stuck, free, free, stuck, free, free, free, stuck, free, free, free, free, …
In this example the joint distribution is variable and tends asymptotically to (1). The process is not stationary.

Stationary process so defined are called **strongly stationary**.

The values of a time series can be summarized by the **mean** a and the **standard deviation** d :
$a=\Sigma_{j=1}^{m} x_j \qquad d=\sqrt{\Sigma_{j=1}^{m}(y_j-a)^2/m}$

For two series of values corresponding to two different random variables:
$y_1, y_2, ..., y_m$ with mean value a
$x_1, x_2, ..., x_m$ with mean value b
it is important to see the **correlation** between them. The intuitive idea of correlation is that two series of values are correlated if they are approximately proportional. Their values grow and decrease more or less at the same time. If they are opposed, when one increase the other decreased it is said that they are negatively correlated. If there is no one of these tendencies they are said uncorrelated. A measure of the correlation is the **correlation coefficient**:

$$\rho = \frac{\Sigma_{j=1}^{m}(x_j - a)\times(y_j - b)}{(\Sigma_{j=1}^{m}(x_j - a)^2 \times \Sigma_{j=1}^{m}(y_j - b)^2)^{1/2}} = \frac{m\Sigma_{j=1}^{m}x_j y_j - \Sigma_{j=1}^{m}x_j \Sigma_{j=1}^{m}y_j}{\sqrt{(\Sigma_{j=1}^{m}x_j^2 - (\Sigma_{j=1}^{m}x_j)^2)\times(\Sigma_{j=1}^{m}y_j^2 - (\Sigma_{j=1}^{m}y_j)^2)}}$$

it can have some real number from –1 (perfect anti-correlation) to 0 (no correlation) to +1 (perfect correlation)

Example 9.2.5.
The two series:
4, 5, 2, 2, 1, 1, 2, 0, 1, 3, 6, 6, 9, 8, 9, 7, 5, 4, 2, 3 mean: 4
8, 3, 6, 4, 3, 4, 4, 5, 4, 7, 9,10,13,10,11,13, 10, 5, 6, 5 mean: 7
are positively correlated. The values centered at the mean are:
0, 1, -2, -2, -3, -3,-2, -4,-3,-1, 2, 2, 5, 4, 5, 3, 2, 0, -2, -1
1,-4,-1, -3, -4, -3,-3, -2, -3, 0, 2, 3, 6, 3, 4, 3, -3, -2, -1, -2
The series:
4, 5, 2, 2, 1, 1, 2, 0, 1, 3, 6, 6, 9, 8, 9, 7, 5, 4, 2, 3 mean: 4
6, 7, 7, 8, 10, 8, 9, 12, 9, 7, 4, 3, 1, 3, 1, 2, 4, 7, 8, 4 mean: 6
are negatively correlated. The values centered at the mean are:

0, 1, -2, -2, -3, -3,-2, -4,-3,-1, 2, 2, 5, 4, 5, 3, 2, 0, -2, -1
0, 1, 1, 2, 4, 2, 3, 6, 3, 1,-2, -3, -5, -3, -5, -4,-2, 1, 2, -2
The series:
4, 5, 2, 2, 1, 1, 2, 0, 1, 3, 6, 6, 9, 8, 9, 7, 5, 4, 2, 3 mean: 4
10,7, 5, 10, 6, 5, 11, 9, 6, 3, 4, 6, 8, 9, 6, 5, 8, 5, 9, 8 mean: 7
have low correlation. The values centered at the mean are:
0, 1, -2, -2, -3, -3,-2, -4,-3,-1, 2, 2, 5, 4, 5, 3, 2, 0, -2, -1
3, 0, -2, 3, -1, -2, 4, 2,-1,-4,-3, -1, 1, 2, -1, -2, 1, -2, 2, 1

note that, in the mean centered values, when the pairs x_i, y_i with equal sign predominated the correlation is positive. When the pairs with different sign predominated the correlation is negative. See

Another important characteristic of a series is the **auto-correlation** of the values. The auto-correlation is the correlation of the values of the series with the values of the same series shifted k places to the left and eliminating the first k values. k is called the **order** of the auto-correlation.

Example 9.2.6. In the previous example the mean and the deviation are:
For the case of the dice: $m = 3.8 \quad d = 1.6$

For the case of the queue $m = 3.4 \quad d = 1.743$
the auto-correlation of first order is between the series :
6,3,2,3,6,4,3,5,5,1,3,6,3,5,2,6,3,5,4.
3,2,3,6,4,3,5,5,1,3,6,3,5,2,6,3,5,4,1
A third order auto-correlation is between the series:
6,3,2,3,6,4,3,5,5,1,3,6,3,5,2,6,3.
3,6,4,3,5,5,1,3,6,3,5,2,6,3,5,4,1.

The measure of auto-correlation is the **auto-correlation coefficient**. For the order k :

$$\rho_k = \frac{m\Sigma_{j=1}^{m-k} y_j y_{j+k} - \Sigma_{j=1}^{m-k} y_j \Sigma_{j=1}^{m-k} y_{j+k}}{\sqrt{(\Sigma_{j=1}^{m-k} y_j^2 - (\Sigma_{j=1}^{m-k} y_j)^2) \times (\Sigma_{j=1}^{m-k} y_{j+k}^2 - (\Sigma_{j=1}^{m-k} y_j)^2)}}$$

Example 9.2.7. For the example of the dice it is obtained (exercise 9.2.2):
$\rho_1 = -0.3217 \quad \rho_2 = -0.2774 \quad \rho_3 = -0.0321$
For the example of the queues:
$\rho_1 = 0.7066 \quad \rho_2 = 0.3890 \quad \rho_3 = -0.011$
in general, the values of queues are positively correlated for low values of k , because the queues fluctuate smoothly: to a large value follows a large value, to a small value follows a small value.
The auto-correlation of order 0 is the variance d^2 .
The results of the simulation are series that are usually correlated (positively or negatively) among them and they are also auto-correlated.

Weak and covariance stationary process. The strong stationarity is difficult to test. An stochastic process is called weakly stationary if for its random variables the mean is constant and the auto-correlations depends only on the order k (and not of the time in which the observations are starting). That is **mean and auto-correlations remain constant in the time**. In particular, the variance remains constant. If, beside this, the correlation between y_i and y_j depends only on the separation j - i (lag) but not on the values of i and j then the process is called **covariance stationary**. Usually the value of the correlation decreases with the separation. Exercise
In the following only covariance stationary process will be considered.

9.2.2 Terminating systems.
In some systems the time of a process depends on properties of the simulated system. They are called terminating systems. In the simulation of terminating systems the simulation time is usually the natural period of the process. It may be determined by a certain value of the time or by the happening of a determined event.

Example 9.2.2.1. In a bank the duration of the operation (and the run of the corresponding model) is given by the time of opening and closing. Or more exactly, the time between the first client arrival and the finishing of service of the last client (before of after the closing). In a simulation of a game between two players the run is finished when one of them retires or get out of money or credit.

In some cases (as in the bank example) the process, from starting to termination, may be repeated but it may be assumed that in the simulation the repetitions are independent processes.
Some systems work without stopping. The non terminating systems may have constant, periodic or stochastic inputs or parameters.

Example 9.2.2.2. In a busy port the rate of entrance of ships have little change in the time. In a traffic light the arrival of cars have some daily fluctuations. The two systems work without stopping.
In economic and ecological systems an annual cycle of production is considered, but although the interactions may be the same each year (and in the models the same algorithm is used each year) the initial conditions are different (may be the end conditions of the previous period)

The simulation of these systems, on the other hand, must start and stop at certain times. Thus, the problem arises of how long the simulation must be to give a useful representation of the system behavior.
In this section terminating system will be considered. Non terminating systems are discussed in 9.2.3.

a) Mean values and confidence intervals for one output variable.
The n **replications** are made using **different random numbers**. It is convenient, for further use in comparisons, that several independent sections of the model use different streams, but this is not essential in the actual discussion of extracting results from runs of only one fixed model. In practice, the replications with different seeds may be done repeating the run without reinitialization of the seeds, i.e. each run uses as seeds the final values of the random numbers in each stream. Other method is initializing explicitly the seeds at different values in each replication. It is practically impossible (in the runs without re-initialization) that in one run the initial seeds were the same than those in other replication, and some correlation is highly unlikely. In this conditions **the different replications may be safely considered independent stochastic processes**.
Let us consider a variable of interest in the real system. In a run we obtain value of this variable. In n replications of the run the values x_i i=1,2,…,n are obtained. These x_i are then independent and, as they are obtained from the same stochastic process, they will have the same distribution.
It is possible to find the **mean value** of the n replications:

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

as a good approximation for the mean μ of the population of **all possible runs**,
and the **standard deviation of the sample**:

$$d = \sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$$

as an estimator of **the deviation** σ of all the runs of size n, assuming that the runs are independent and therefore the sequence of the x_i has not autocorrelation.
If **all the possible samples** of size n were taking , it can be shown (because the x_i are independent and have the same distribution) that **the values** $\frac{\bar{x}-\mu}{\sqrt{d^2/n}}$ **have a Student** t **distribution with** $n-1$ **degrees of freedom**, where μ is the mean of the population of all possible samples of size n.

This result allows to test hypotheses about means.

For **one sample of n replications** the interval with a confidence level of $1-\alpha$ for $\bar{x}$ is given by:

$\bar{x} \pm t_{n-1}(1-\alpha/2)\sqrt{d^2/n}$

where $t_{n-1}(1-\alpha/2)$ is the value of t for which the Student *t* distribution with $n-1$ degrees of freedom has the value $1-\alpha/2$. The meaning of the confidence interval is the following: if the population of all the intervals estimated in this way (obtained from all the possible samples of size n) is considered, the probability that they contain μ is greater than $1-\alpha$. That is a fraction $1-\alpha$ or more of them contain the true mean.
The size of this confidence interval gives an estimation of the statistical error of the $\bar{x}$ obtained in n replications with a prefixed level of confidence α.
See that the interval size increases proportionally to d, and decreases as $\sqrt{n}$. Multiplying the number of replications by four the size of the interval decrease only to a half (a little more because z_{n-1} decreases with n). As more confidence is sought the interval increases, because $z_{n-1}(1-\alpha/2)$ increases when $1-\alpha/2$ increases.

Example 9.2.2.4. A model of a simple queue with arrival times 4.0 and serving time 3.8 both from an exponential distribution, was simulated for a time of 10000. Using different random numbers. The results of 10 replications for the mean delay in the queue were found:

```
171.892, 77.520, 39.394, 51.034, 51.500, 43.597, 81.950, 23.987
142.382, 59.092
```

The mean and deviation are: $\bar{x} = 74.2352 \quad d = 47.3919$
From the table for the values of $t(z)$ for $\alpha = 0.05,\ 1-\alpha = 0.95$ and 10-1=9 degrees of freedom we obtain $t_{n-1}(1-\alpha/2)$ =1.833; and:

$\bar{x} \pm t_{n-1}(1-\alpha/2)\sqrt{d^2/n}$ =74.2352 ± 27.4705

The, μ is in the intervals found like the (46.7647, 101.7057) with a probability of 0.95. See that the error is rather high.

b) Number of runs to get a given error e with a certain confidence level.
An approximated method is to start from a confidence interval with a tentative n (usually low) and use:

$$n_\alpha(e) = MIN(i : t_{i-1}(1-\alpha/2)\sqrt{d^2/i}) \le e)$$

starting with $i = n+1$ and assuming that d does not change with i, the minimum *i* for which the inequality holds is reached increasing i one by one or looking by trial and error.

9.2.2.5. To get an interval of 2×10 (instead of the previous 2×27.4705 seen in a)) by this method the size i of the sample (number of replications) must be such that:

$t_{n-1}(1-0.05)\sqrt{47.3919^2/i} < 10$ (1)

it is seen by trial and error that i must be less than 75 so the value $t_{75}(1-0.05/2)=1.665$ may be safely used. Then:

$$i \geq \frac{1.665^2 \times 47.3919^2}{10^2} = 62.26$$

so 63 replications are enough.

The method assume that d do not change with n which may be a questionable assumption. More exact is to do a sequential experiment: at each step, starting from the original sample, a new sample is obtained and the new d is estimated and with this d the (1) is checked, adding replications until the inequality is fulfilled (see Exercise 9.2.4). Note that now, i will depend of the results of these additional replications. With different random numbers a different i may be obtained.

9.2.3 Non terminating experiments. Steady state.

Some non-stop systems works in a steady state (as in the case of a port mentioned above). The simulation usually starts from conditions that do not correspond to this steady state, because almost always the condition of the steady state are not known in general o for a particular run. If the model contains the appropriate feed-backs, **a steady state will be reached after a time of transient behavior**. This steady behavior is the interesting to the user. So the first problem is to find when the transient period finish and the steady state starts. **Only the statistics for the steady state must be considered.**

Example 9.2.3.1. In a queue persons arrive with times from an exponential distribution with mean 4min. They are served in times from a gamma distribution with mean 7min. and deviation 2min.. If the queue is greater than 15, 90% of the people do not enter and leave the system.
For a run of length 700min the following values are obtained:
length of the queue: mean: 14.940 deviation: 3.392
waiting time in the queue mean: 80.632 deviation: 44.520
The large deviations are caused for a transient from the void initial state to the steady regime. From the graphic of the time series for the queue, it is estimated that the transient is about 120min.
If the statistics are counted from 120 to 700 the following values are obtained:
length of the queue: mean: 15.281 deviation: 1.166
waiting time in the queue mean: 105.289.deviation: 11.687
These values have less deviation and represent more exactly the steady state. Longer runs do not substantially improve this result. In a run from 120 to 10000 the values were:
length of the queue: mean: 15.320 deviation: 0.978
waiting time in the queue mean: 108.537.deviation: 10.470
On the other hand, in the runs without elimination of the transient, the effect of this remains for a long time. For example in a run from 0 to 2000 the values were:
length of the queue: mean: 14.281 deviation: 2.189
waiting time in the queue mean: 95.217.deviation: 30.338
that is worse than the short run 120 to 700.
See Exercise 9.2.5..

The elimination of the transient gives better results with shorter simulation runs.

The first idea is to start the simulation with the conditions of the steady state. So, in a system with queues and resources, instead of starting with the system void, to start with the length of the queues and occupation of the resources corresponding to the steady state. These values are, of course, not known (they are one of the targets of the model building) so that only a guess of these values may be assign. See the example 7.2.1.

Methods to find when the steady state begins.
It is not always clear in what point in the time the steady state starts. The approaching to the steady state may be gradual and, in stochastic models, concealed by strong fluctuations.
The inspection of the graphic of the output is, of course, the more easy method and it works in many cases. For other cases many methods are proposed.
a) **Values above and below the mean**. Consider the sequence of values $x_1, x_2, x_3, ..., x_n$ of a run. There is a steady state from the k result onwards if the mean $\bar{x}_k$ of $x_{k+1}, x_{k+2}, x_{k+3}, ..., x_n$ leaves about the same quantity of terms of the sequence above and below.

Example 9.2.3.2. The following values are obtained in a stochastic model of growth:

```
0.701 1.153  1.614  1.987  2.257  2.495  2.687  2.813  2.913  2.95
3.910 3.202  3.225  3.345  3.331  3.344  3.415  3.398  3.415  3.452
3.441 3.434  3.389  3.539  3.448  3.517  3.543  3.443  3.568  3.461
3.452 3.563  3.509  3.461  3.508  3.420  3.427  3.461  3.444  3.513
```

the following means are obtained:

$\bar{x}_0 = 3.170$ Points below: 10 Points above: 29
$\bar{x}_{10} = 3.425$ Points below: 11 Points above: 19
$\bar{x}_{15} = 3.463$ Points below: 8 Points above: 17
$\bar{x}_{20} = 3.477$ Points below: 12 Points above: 8
$\bar{x}_{24} = 3.493$ Points below: 7 Points above: 9
After 24 values the difference between the number of points above and below seems to be due to statistical fluctuations, so, from 3.448 onwards the results appears to be steady with a mean of 3.493.
The exact mean of the theoretical model is 3.5.
Although this method is too simple and some counter examples can be found, it works in many cases and is a good aid to the graphical method.

b) **Moving average method**. A moving average is a series of averages of k successive values starting from the 1st, then for 2nd, then for 3rd, etc. That is the following means are computed:
$(x_1 + x_2 ... + x_k) / k$, $(x_2 + x_3 + ... + x_{k+1}) / k$,
$(x_3 + x_4 + ... + x_{k+2}) / k$,.., $(x_{n-k} + x_{n-k+1} + ... + x_n) / k$
When these averages do not change significantly the steady state is assumed to be reached.

Example 9.2.3.3. In the above example, taken the averages of size 5, that is, doing the successive means of
```
0.701  1.153  1.614  1.987  2.257 ;
```

```
.  . . .  . .. .    . . . . . . .
2.495  2.687  2.813  2.913  2.957
. . . . . . . . . . . . . . . . .
3.420  3.427  3.461  3.444  3.513
```
the following moving average are found:
```
1.542  1.901  2.208  2.448  2.633  2.773  2.892  2.995  3.077  3.164  3.239
3.289  3.332  3.367  3.381  3.405  3.424  3.428  3.426  3.451  3.450  3.465
3.487  3.498  3.504  3.507  3.494  3.498  3.511  3.489  3.499  3.492  3.465
3.456  3.452  3.453
```
It may be observed that after the value 3.451 there are greater and lower values of the moving average. This is the mean of the group: $x_{20}, x_{21}, x_{22}, x_{23}, x_{24}$. So after x_{24} the values enter in a steady series. One problem is to define the size of k. The more fluctuating is the series the larger must be k to smooth the fluctuations.

Mean estimation from the steady state. Once the steady state is define the problem is to estimate the mean value from the time series from which the transient was excluded. Of course the best estimation of the μ corresponding to an infinite run is the mean of the values. The way to estimate the mean depend of the type of time values considered. If it is a discrete set $t_1, t_2, ..., t_n$ as in the models with fixed interval time, the mean value is computed from the values at each time:

$\bar{x} = \Sigma_{i=1}^{n} x_i / n$

If the time is a continuous variable

$$\bar{x} = \frac{1}{t_{sim} - t_0} \int_{t_0}^{t:=t_{sim}} x dt$$

where t_0 is the time in which the steady state begins and t_{sim} is the time of the simulation (or the maximum time for which the values of x are to be considered).

Example 9.2.3.4. In a queue the following changes of the length are observed after the transient:

Times	23	27	28	32	45	47	56	61	66	70	77	79	81
Length	9	10	11	12	11	10	11	10	9	8	7	8	9

The queue has defined values for all the real values of the time (for example for the time 25 it is 9). So the integral formula must be applied to find the exact mean:

$\bar{x}$ =((27-23)×9+(28-27)×10+(32-28)×11+….+(81-79)×8)/(81-23)=10.0862

If the only data are observations of the queue each 5 seconds starting at 23 the defined data are:

Lengths 9 10 12 12 12 10 10 11 10 9 8 7

Whose mean is: (9+10+12+12+12+10+10+11+10+9+8+7)/12=10.0000

What is an approximate value to the true mean.

The variance (square of deviation) of a set of data is estimated by :

$d^2 = \Sigma_{i=1}^{n} (x_i - \bar{x})^2 / n$

in the discrete case, and by:

$$d^2 = \frac{1}{t_{sim} - t_0} \int_{t_0}^{t:=t_{sim}} (x - \bar{x})^2 dt$$

in the continuous time.

It is seen in Statistics that the best estimation of the variance σ^2 of the population is not the variance of the population but the expression:

$d^2 = \Sigma_{i=1}^{n}(x_i - \bar{x})^2 /(n-1)$

This estimation is not biased if the series of data do not have auto-correlation, that is to say each value is not dependent from the previous values. But this is not usually true for simulation results. In a que ue, for instance to a great value follows also a great one.

Example 9.2.3.5. The data of a queue in a model are taken at each unit time. The model is a simple queue with exponential times for arrivals (mean 2) and services (mean 1.9) so that the mean length of the queue (see) is 1.9/(2-1.9)1=19. A succession of 200 results taken each second in the steady state are:

```
 7, 7, 7, 7, 7, 7, 7, 8, 8, 9,
 9, 8, 8, 8, 8, 8, 6, 3, 4, 5,
 6, 7, 7, 6, 5, 3, 3, 0, 0, 0,
 0, 1, 2, 3, 6, 6, 8, 9, 9,10,
10,10,11,12,14,13,14,15,16,15,
15,14,15,16,17,16,17,17,18,18,
19,18,17,18,17,18,18,19,20,21,
21,22,22,23,22,21,22,24,24,26,
25,25,27,29,25,22,20,20,19,17,
15,14,13,13,12,12,13,13,11,10,
10, 9, 8, 8, 8, 9, 8, 7, 6, 6,
 6, 6, 6, 5, 6, 6, 7, 7, 7, 6,
 5, 5, 4, 5, 5, 5, 5, 5, 5, 0,
 0, 0, 2, 3, 2, 1, 2, 1, 2, 3,
 1, 3, 3, 4, 5, 5, 5, 4, 4, 5,
 5, 7, 6, 6, 6, 7, 7,10,10,11,
12,12,11,12,14,14,16,18,19,21,
19,17,17,15,13,11, 9, 7, 9, 9,
10,10, 7, 7, 7, 5, 6, 6, 7, 8,
 8, 8, 9,10, 9, 6, 6, 6, 7, 6
```

The mean is 10.15, that is the best, unbiased estimation of the population mean, and the estimation of the population of the deviation, computed by the usual form is $d = 6.466$. So an estimation of the deviation of the mean is $10.15/\sqrt{6.466}$ =3.9916.

The comparison with the theoretical value of the mean (19) shows that this estimation of the deviation is not correct. It must be greater. It can be shown (see Anderson 1994) that an unbiased estimation of the deviation of the mean d_u is given by:

$$d_u^2 = \frac{d^2}{n}\left(1 + 2\sum_{k=1}^{m}(1 - \frac{k}{m+1})\rho_k\right)$$

where ρ_k is the auto-correlation of order k of the series (see 9.2). The sum might be extended indefinitely. For practical reasons, the sample size is finite and m is limited by the sample size. Normally the correlation coefficients diminishes as k increases and for large values of *k* the signs alternate and the terms may cancel each other. As a rule of thumb (see Shannon 1975) m may be taken about 10% of the sample size.

An estimation of the sample size when auto-correlation exist is given by:

$$n = \frac{(t_{\alpha/2})^2 s^2 \{1 + 2\sum_{p=1}^{m}(1 - \frac{p}{m+1})\rho_p}{d^2 \bar{x}^2}$$

where *d* is the desired precision, i.e. with this value of n the estimated mean $\bar{x}$ will differ from the real mean in less than *d* (by defect or excess) with a confidence level of α.

EXERCISES

Exercise 9.2.1. Compute the matrix of frequencies for the example 9.2.2. Show that the process of successive selections of colour (red or blue) and number (0 or 1) is a stationary stochastic process. Show that the sequences of each variable are stochastic processes.
Show that the variables are correlated (not independent).

Exercise 9.2.2. Find the correlation coefficients of the examples 9.2.4.

Exercise 9.2.3. Give an argument to see that in a stable queue the autocorrelation coefficients decreases with the order of autocorrelation.

Exercise 9.2.4. Program Example 9.2.10 an estimate the number of replications for a confidence interval of (-10,10) considering the change in *d*.

Exercise 9.2.5. Program example 9.2.12 and see the graph of the queue. Compare with the estimated transient.

Exercise 9.2.6. Use the data given in Example 9.2.3.5 to otain a sample size with *d*=0.1 at a confidence level of 0.05.

BIBLIOGRAPHY

Zeigler B. Theory of Modeling and Simulation. P.131. Wiley, 1975.