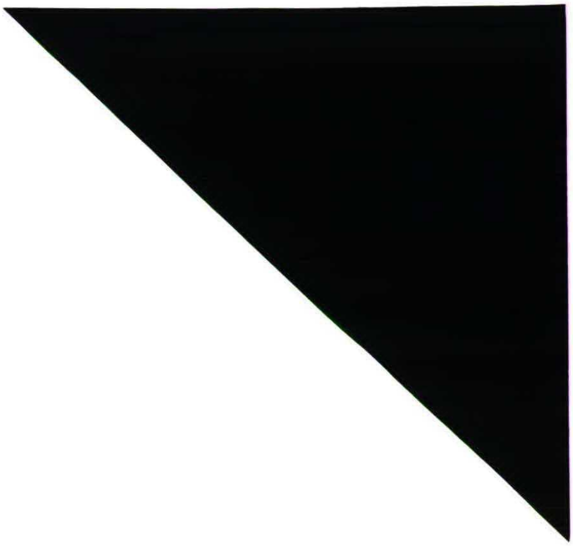
Research Memorandum

Faculty of Economics and Business Administration

Tilburg University



New proposals for the validation of trace-driven simulations

J.J.A. Moors and
L.W.G. Strijbosch

FEW 745

Communicated by Prof.dr. J.P.C. Kleijnen

NEW PROPOSALS FOR THE VALIDATION OF TRACE-DRIVEN SIMULATIONS

J.J.A. Moors and L.W.G. Strijbosch

Abstract

Simulations try to mimick reality. Hence, it is desirable that simulated values resemble the values observed in real life. Although the problem how to validate a simulation model has received quite a lot of attention, even recent literature offers no generally accepted standard method.

This paper proposes a general approach. For trace-driven simulation, three different objectives are formulated. At the one extreme, the individual outcomes of a real life process and the simulated process should agree. An intermediate objective is that the outcomes of the two processes should have a symmetric joint distribution. And at the other extreme, only the marginal distributions of both outcomes have to be comparable. For all three objectives, validation measures and/or tests are proposed. For the last objective, a simulation study is presented, comparing two competing tests.

Keywords: Kleijnen test, mimicking simulations, simulation objectives, Stuart test, trace-driven simulations, validation.

1 Introduction

General discussions on validation of simulation models can be found in all textbooks on simulation, e.g., LAW & KELTON (1991, p. 298-324) and PEGDEN et al. (1990, p. 133-162). Recent survey papers on this subject are BALCI (1994) and KLEIJNEN (1995), including 102 and 61 references, respectively. The extensive literature, however, does not offer a standard theory on validation, nor a standard criterion to measure the quality of simulation models.

This paper will differentiate between the following two types of simulation experiments; attention will be focussed on the first type.

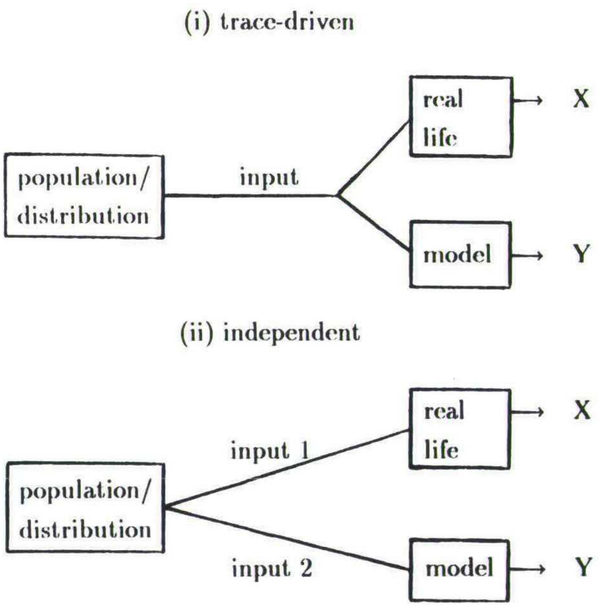
(i) Each simulation is related to precisely one real life situation, because the same input values are used; hence, the simulated value should be close to the observed real life value. Simulationists call this trace-driven simulation.

(ii) No one-to-one relation exists between the simulation experiments on the one hand and the real life situations on the other. Interest centers on the distribution of the simulated values and the observed values separately. The simulation is considered successful if these two distributions are similar. We will call this independent simulation.

Both the real life values and the simulated values can be seen as outcomes of a random variable; denote these by X and Y, respectively. Statistically speaking, in trace-driven simulation the joint distribution of the pair (X, Y) is of interest. In independent simulation on the other hand, there is no relation between individual X and Y values: only the marginal distributions of X and Y can be taken into account.

Figure 1 pictures the two types of simulation experiments.

Figure 1. Trace-driven versus independent simulation.



Inputs are generated from either an experimental or a theoretical distribution; 'real life' is seen as a black box that has been modeled by the simulationist. Note that the segregation between the two types is not always complete: data from the black box may be used in the model.

In case (ii), the only possible validation question is: Are the distributions of X and Y identical? To answer this question, a number of well-known tests is available - since the observations on X and Y are independent. Examples are the two-sample Kolmogorov-Smirnov test (also called the Smirnov test) and the χ^2-test for homogeneity; see LINDGREN (1976) or D'AGOSTINO & STEPHENS (1986). Note that the corresponding test statistics can be used as validation measures. Since case (ii)allows numerous well-known answers, only case (i) will be considered henceforth.

The most recent attempt to formulate validation criteria for case (i) – trace-driven simulation – was offered by KLEIJNEN et. al. (1997). They assumed a bivariate normal model for (X, Y):

$$(X,Y) \sim N_2 \left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right]$$

and proposed the following 'stringent validation requirement':

$$\mu_x = \mu_y = \mu, \quad \sigma_x^2 = \sigma_y^2 = \sigma^2, \quad \rho > 0 \tag{1.1}$$

Arguing that in most trace-driven simulations ρ will be positive, they do not test the last part of (1.1). Their validation procedure therefore consists of testing the joint hypothesis

$$H_0 : \mu_x = \mu_y, \sigma_x^2 = \sigma_y^2 \tag{1.2}$$

A simulation model is considered satisfactory, if H_0 is not rejected. The new test they developed is explained in some more detail in Section 4.

In our view, the Kleijnen validation procedure described above is very satisfactory - within the limitation of their assumptions. We think, however, that these assumptions are too restrictive to make their procedure generally applicable. To make this argument more precise, three main objections are listed below.

1) The requirement on ρ in (1.1) is rather weak. In general, the purpose of trace-driven simulation is to achieve a **high** (positive) correlation between X and Y. Further, we fail to see why a model with, say, $\rho = 0.1$ is satisfactory, whereas $\rho = -0.1$ is not.

2) In the subsequent testing problem, ρ does not take part: instead, $\rho > 0$ is assumed to hold. If a positive ρ is a 'stringent validation requirement', the validation procedure should test this assumption.

3) While in fact equality of the marginal distributions of X and Y is of interest, the Kleijnen procedure is only concerned with equality of means and variances. Of course, if indeed the binormality assumption is satisfied, (1.2) is equivalent to equality of the two marginal distributions. If not, however, identical means and variances can occur for quite different distributions.

In view of these objections, new validation measures will be proposed for trace-driven simulation. Our measures depend on the purpose of the simulation experiment; three different cases will be treated.

A) In general, the objective of the simulation experiment will be to mimick a real process or phenomenon. We will call this kind of experiment (real life) mimicking simulation. The ideal, utopian situation is that each simulated value exactly equals the observed real life value: $X = Y$.

B) At an intermediate level, the simulation experiment can be considered satisfactory, if the joint distribution of (X, Y) is symmetric. In the terminology of DE FINETTI (1974), this means that X and Y are exchangeable.

C) Finally, even if trace-driven simulation is used, interest may be centered on the marginal distributions of X and Y only. Now the earlier validation question returns: Are the distributions of X and Y identical? Note, however, that X and Y are dependent now.

Observe the following chain of implications:

$$A) \Rightarrow B) \Rightarrow C).$$

A validation measure for the most important case A) - mimicking simulation - is presented in Section 2; an estimator for this measure is also discussed. Section 3 treats cases B) and C); in Section 4 the test proposed for case C) is compared with the test suggested by KLEIJNEN et al. (1997), by means of a simulation study. The final Section 5 gives some conclusions and discussion.

2 A validation measure for mimicking simulations

In any trace-driven simulation experiment, the ideal situation is that each simulated value exactly equals the observed value. In this utopian case $Y = X$ holds or, equivalently,

$$\mu_y = \mu_x, \quad \sigma_y = \sigma_x, \quad \rho = 1 \tag{2.1}$$

The proposed measure is based on the deviations from this utopy, that is, on the differences $\mu_y - \mu_x$, $\sigma_y - \sigma_x$, and $1 - \rho$. More specifically, we propose as ***validation measure*** M for real life mimicking simulations (case A):

$$M = [(\mu_y - \mu_x)^2 + (\sigma_y - \sigma_x)^2 + 2(1-\rho)\sigma_x\sigma_y]/\sigma_x^2 \tag{2.2}$$

Nine advantages and interesting features of M are listed below.

a) All three terms in the numerator are non-negative; hence, lower values of M can be attained only by moving into the general direction of the utopian situation (2.1).

b) Positive and negative differences $\mu_y - \mu_x$ and $\sigma_y - \sigma_x$ are treated identically, because of the squares in the numerator. This agrees more or less with intuition, although $\sigma_y < \sigma_x$ may tend to occur more frequently in practice, because minor causes of variability might not be represented in the model.

c) M is dimensionless: X and Y have the same dimension, and the same holds for all three terms in the numerator.

d) M is location-scale invariant: the simultaneous transformations $X^* = aX + b, Y^* = aY + b$ leave M unchanged.

e) Define $D = Y - X$ with mean μ_d and variance σ_d^2. Because of the relation

$$(\sigma_y - \sigma_x)^2 + 2(1-\rho)\sigma_x\sigma_y = \sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2 = \sigma_d^2$$

our validation measure may be written as

$$M = (\mu_d^2 + \sigma_d^2)/\sigma_x^2 = E(D^2)/\sigma_x^2 \tag{2.3}$$

In words, M measures the expectation of the squared deviation $Y - X$ in units σ_x^2. The smaller this deviation is on average, the better the simulation experiment is; this is intuitively appealing.

f) Introducing

$$\nu = \mu/\sigma, \tau = \sigma_y/\sigma_x$$

(2.2) can be rewritten again as

$$M = (\tau\nu_y - \nu_x)^2 + (1-\tau)^2 + 2\tau(1-\rho) \tag{2.4}$$

So M depends only on four parameters, viz. the coefficients of variation ν_x and ν_y, the relative standard deviation τ, and the correlation coefficient ρ.

g) The optimal value $M = 0$ implies the utopian situation (2.1), or $X = Y$. The practical implication is that for low M-values, the joint distribution of (X, Y) is almost degenerated. Hence, in this case it is not necessary to check the objectives B) and C).

Admittedly, criterion (2.2) shows some arbitrary elements; in particular, the choice of equal weights in the numerator may be questioned. In particular, differences between means might be considered to be more serious than differences between the two standard deviations. However, giving $(\mu_y - \mu_x)^2$ a larger weight than $(\sigma_y - \sigma_x)^2$ would imply the loss of the attractive property e).

Estimators for M are readily available. Let n denote the number of i.i.d. replications, i.e., the number of observed pairs (X_i, Y_i). If σ_x^2 were known, (2.3) would lead to the unbiased estimator

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n} D_i^2/\sigma_x^2 \tag{2.5}$$

for M. For unknown σ_x^2, a straightforward extension is

$$\tilde{M} = \frac{1}{n}\sum_{i=1}^{n} D_i^2/S_x^2 \tag{2.6}$$

with $S_x^2 = \sum^n (X_i - X)^2/n$. (Note that $\tilde{M}$ will be asymptotically unbiased.)

Some properties of the (common) numerator $T = \frac{1}{n}\sum_{i=1}^{n} D_i^2$ in (2.5) and (2.6) are presented now. Denote for any random variable X

$$\mu = E(X); \quad \mu_k = E(X-\mu)^k, \quad k = 2, 3, 4, \ldots.$$

Then well-known results are

$$\left.\begin{aligned} E(X^2) &= \mu_2 + \mu^2 \\ E(X^4) &= \mu_4 + 4\mu\mu_3 + 6\mu^2\mu_2 + \mu^4 \\ V(X^2) &= \mu_4 + 4\mu\mu_3 + 4\mu^2\mu_2 - \mu_2^2 \end{aligned}\right\} \tag{2.7}$$

The variance of D^2 follows, as well as the variance of T. The Central Limit Theorem gives that T is approximately normally distributed.

If (X, Y) has the binormal distribution, (2.7) reduces to

$$V(X^2) = 2\mu_2^2 + 4\mu^2\mu_2$$

so that

$$V(D^2) = 2\sigma_d^2(\sigma_d^2 + 2\mu_d^2)$$

Consequently

$$T \approx N\left[\mu_d^2 + \sigma_d^2,\ 2\sigma_d^2\left(\sigma_d^2 + 2\mu_d^2\right)/n\right]$$

where $\approx$ means 'approximately distributed as'.

3 Tests for symmetry and homogeneity

In this section, tests (and measures) for the cases B) and C) will be considered, based on χ^2-distributions. Split the range of possible values of X and Y into c classes A_i; denote the joint probability $P(X \in A_i,\ Y \in A_j)$ by p_{ij}. See Table 1.

Table 1. Cross-classification of (X, Y)

y / x	A_1	A_2	$\cdots$	A_c	Total
A_1	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_c	p_{c1}	p_{c2}	$\cdots$	p_{cc}	$p_{c\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot c}$	1

The corresponding (random) frequencies of the n replications of the trace-driven simulation experiment are denoted by N_{ij}, $N_{i\cdot}$ and $N_{\cdot j}$.

Case B) considers the question of exchangeability; if X and Y are exchangeable, Table 1 should be symmetric, so that the testing problem is

$$H_0 : p_{ij} = p_{ji}, \text{ for all } i \neq j$$
$$H_1 : p_{ij} \neq p_{ji}, \text{ for some pair } (i, j).$$

An approximate χ^2-test for this problem is easily derived. The unconditional Maximum Likelihood (ML) estimators for the p_{ij} are N_{ij}/n, whereas under H_0 the ML-estimators are

$$N_{ii}/n, \quad (N_{ij} + N_{ji})/(2n)$$

By consequence, the well-known χ^2-statistic, say G, becomes

$$G = \sum_{i} \sum_{\substack{j \\ i \neq j}} \frac{[N_{ij} - (N_{ij} + N_{ji})/2]^2}{(N_{ij} + N_{ji})/2}$$

which may be simplified to

$$G = \sum_{i} \sum_{\substack{j \\ i < j}} \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}} \tag{3.1}$$

Under H_0, the approximate distribution of G is given by

$$G \approx \chi^2_{c(c-1)/2} \tag{3.2}$$

Note that for $c = 2$, this test reduces to the well-known McNemar test; see McNEMAR (1947).

The testing problem in case C) - homogeneity - can be formulated as

$$H_0 : p_{i\cdot} = p_{\cdot i} \text{ for all } i$$
$$H_1 : p_{i\cdot} \neq p_{\cdot i} \text{ for some } i$$

The dependence between X and Y implies that the standard χ^2-test for homogeneity does not apply. Since the ML-equations appear to be intractable in this case, the problem is more complicated. STUART (1955) gave a solution. Since his paper seems to be relatively unknown, we present a more detailed description of the Stuart test.

Natural starting point in testing H_0 are the statistics

$$D_i = N_{i\cdot} - N_{\cdot i}, \quad i = 1, 2, \cdots, c$$

with the following expectations and (co)variances:

$$\left.\begin{array}{l} E(D_i) = n(p_{i\cdot} - p_{\cdot i}) \\ V(D_i) = n(p_{i\cdot} + p_{\cdot i} - 2p_{ii}) - n(p_{i\cdot} - p_{\cdot i})^2 \\ C(D_i, D_j) = -n(p_{ij} + p_{ji}) - n(p_{i\cdot} - p_{\cdot i})(p_{j\cdot} - p_{\cdot j}) \end{array}\right\} \tag{3.3}$$

Since $\sum_{i=1}^{c} D_i = 0$, the D_i are linearly dependent and the limiting distribution for $n \to \infty$ of

$$D^T = (D_1, D_2, \cdots, D_{c-1})$$

is a $(c-1)$-dimensional normal distribution. So, under H_0 and for large n, the approximation

$$D \approx N_{c-1}(0, C)$$

holds, where the elements of the covariance-matrix C follow from (3.3):

$$\left.\begin{array}{l} c_{ii} = n(p_{i\cdot} + p_{\cdot i} - 2p_{ii}) \\ c_{ij} = -n(p_{ij} + p_{ji}) \end{array}\right\}$$

Consequently, $D^T C^{-1} D \approx \chi^2_{c-1}$; the same limiting distribution arises when C is replaced by the matrix V of estimated covariances with elements

$$\left.\begin{array}{l} v_{ii} = n_{i\cdot} + n_{\cdot i} - 2n_{ii} \\ v_{ij} = -n_{ij} - n_{ji} \end{array}\right\} \tag{3.4}$$

Hence, Stuart's test statistic G_S is

$$G_S = D^T V^{-1} D \tag{3.5}$$

with $G_S \approx \chi^2_{c-1}$ holding under H_0.

Note that (3.1) and (3.5) can be used not only as (validation) test statistics, but as validation measures as well.

4 Simulation of homogeneity tests

The central question in case C) is: Are the (marginal) distributions of X and Y identical? Two tests were proposed. The distribution-free approach in Section 3 leads to the null hypothesis

$$H_0 : p_{i\cdot} = p_{\cdot i} \text{ for all } i \tag{4.1}$$

with $p_{i\cdot} = P(X \in A_i)$ and $p_{\cdot i} = P(Y \in A_i)$. The Stuart χ^2-test is then applicable.

In KLEIJNEN et al. (1997) the null hypothesis

$$H_0 : \mu_x = \mu_y, \;\; \sigma_x^2 = \sigma_y^2 \tag{4.2}$$

was proposed; under the assumed binormality of (X, Y), this is equivalent to identical marginal distributions. Using the notation

$$D = X - Y, \;\; S = X + Y$$

and the linear regression model

$$D = \gamma_0 + \gamma_1 S + \varepsilon$$

(4.2) is equivalent to

$$H_0 : \gamma_0 = 0, \;\; \gamma_1 = 0 \tag{4.3}$$

This hypothesis can be tested by means of a familiar F-test. Denote the estimated model by

$$\hat{D} = C_0 + C_1 S$$

where Ordinary Least Squares is used to estimate the regression coefficients γ_0 and γ_1. Then the test statistic G_K in the Kleijnen test is

$$G_K = \frac{n-2}{2} \frac{\sum^n D_i^2 - \sum^n (D_i - \hat{D}_i)^2}{\sum^n (D_i - \hat{D}_i)^2} \tag{4.4}$$

Now, $G_K \sim F_{2,n-2}$ holds under null hypothesis (4.3).

To compare the power of these two tests a simulation study was performed. (As a byproduct, this study gives insight into the properties of the relatively unknown Stuart test.) This simulation study consists of two separate parts. In *part A*, the joint distribution of X and Y is given by

$$(X, Y) \sim N_2 \left[\begin{pmatrix} 0 \\ \mu_y \end{pmatrix} \right], \left[\begin{pmatrix} 1 & \rho\sigma_y \\ \rho\sigma_y & \sigma_y^2 \end{pmatrix} \right] \tag{4.5}$$

Hence, the binormality assumption is satisfied. Since this assumption is essential for the Kleijnen test and irrelevant to the Stuart test, the latter is expected to be inferrior in this case. More precisely, for the same level of significance α the Kleijnen test should have higher power. Part A of our simulation experiment was designed to check this.

For different sets of values for the parameters μ_y, σ_y and ρ, 2000 random samples of size 200 were drawn from the joint distribution (4.5). For each sample the outcomes g_K and g_S of the test statistics in (4.4) and (3.5) were calculated. When applying the Stuart test, $c = 10$ equi-probable classes were formed, based on the standard normal. Then, using the notation $\#A$ for the number of occurrences of event A, the following fractions were calculated:

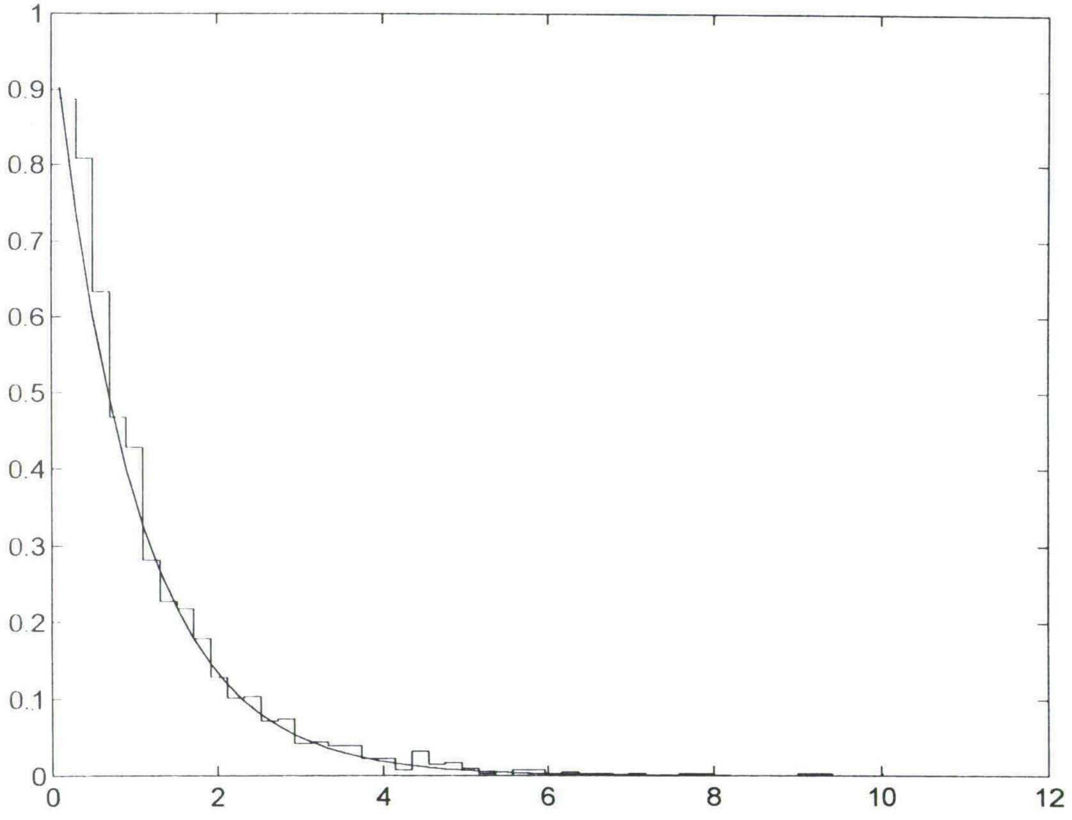
$$\hat{\beta}_K = \frac{\#\{g_K > F_{2,198;\alpha}\}}{2000}, \quad \hat{\beta}_S = \frac{\#\{g_S > \chi^2_{9;\alpha}\}}{2000} \tag{4.6}$$

For the parameter choice $\mu_y = 0, \sigma_y = 1$, the null hypotheses (4.1) and (4.2) are true; then the expectations of the above fractions should equal α. For other values of (μ_y, σ_y), (4.6) estimates the power β of the two tests.

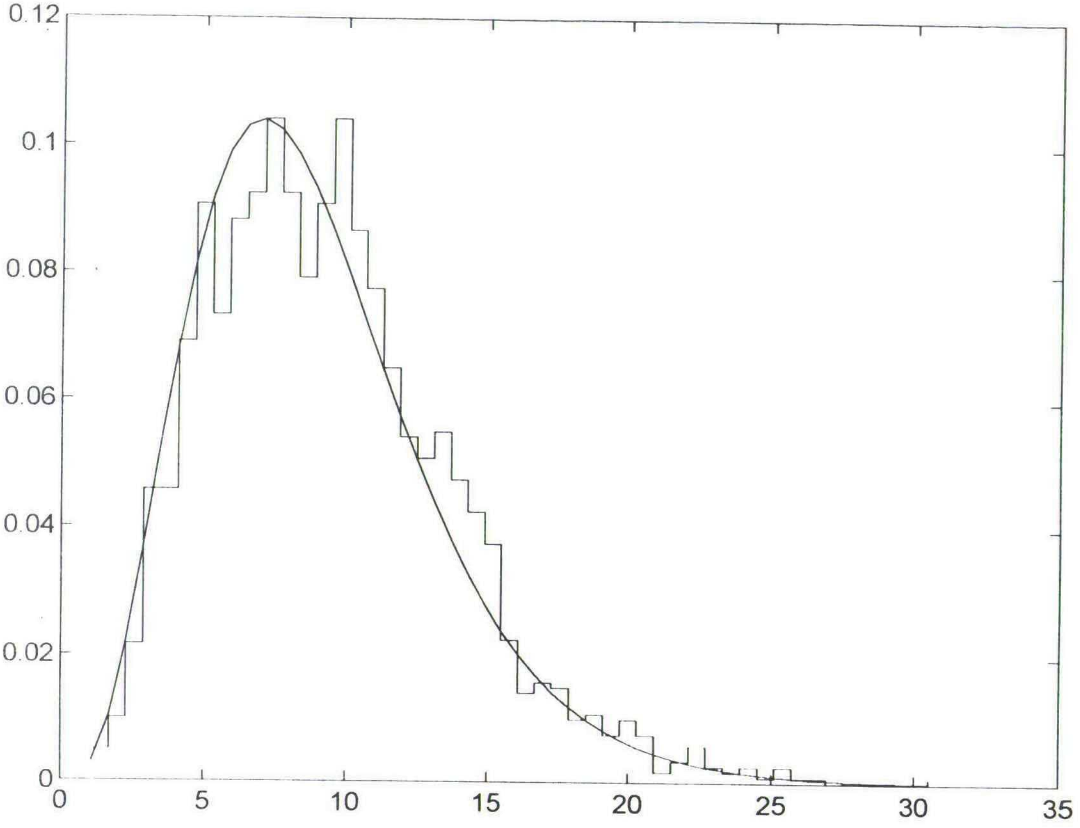
Figure 2 shows a typical picture of the behaviour of G_K and G_S for $\mu_y = 0$ and $\sigma_y = 1$; here $\rho = 0.3$. Figure 2(a) shows the frequency distribution of the (2000) g_K-values, as well as the corresponding theoretical distribution $F_{2,198}$; Figure 2(b) pictures the g_S-values and the χ^2_9-distribution.

Figure 2. Empirical and theoretical distributions of the test statistics G_K and G_S under H_0; normal case with $\rho = 0.3$.

(a) Kleijnen test ($F_{2,198}$)



(b) Stuart test (χ^2_9)



Both empirical distributions are in close agreement with the theoretical ones.

The results of our extensive simulations are summarized in Table 1, which consists of four sub-tables. For different values of the pair (μ_y, σ_y) the simulated probabilities of rejecting H_0 are presented; the numbers between parentheses refer to the Stuart test. Bold faced numbers refer to the size of the test (α). To indicate the numerical precision of the tabulated fractions: the 95%-confidence margin for the true power β increases from 0.010 (for $\hat{\beta} = 0.05$ or 0.95) to 0.022 for $\hat{\beta} = 0.5$. The sub-tables relate to different values of α and ρ.

Table 1. Simulated powers $\hat{\beta}_K$ and ($\hat{\beta}_S$); normal case.

(a) $\alpha = 0.05, \rho = 0$

μ_y \ σ_y	0.9	1.0	1.1	1.2
0	0.24	**0.05**	0.20	0.63
	(0.10)	**(0.05)**	(0.09)	(0.22)
0.1	0.34	0.13	0.31	0.69
	(0.14)	(0.09)	(0.11)	(0.27)
0.2	0.63	0.40	0.56	0.83
	(0.30)	(0.21)	(0.24)	(0.39)

(b) $\alpha = 0.05, \rho = 0.6$

μ_y \ σ_y	0.9	1.0	1.1	1.2
0	0.34	**0.05**	0.30	0.84
	(0.13)	**(0.05)**	(0.11)	(0.31)
0.1	0.59	0.27	0.51	0.89
	(0.23)	(0.14)	(0.18)	(0.39)
0.2	0.94	0.81	0.87	0.97
	(0.64)	(0.48)	(0.49)	(0.65)

(c) $\alpha = 0.1, \rho = 0$

μ_y \ σ_y	0.9	1.0	1.1	1.2
0	0.34	**0.11**	0.32	0.74
	(0.18)	**(0.10)**	(0.15)	(0.35)
0.1	0.47	0.21	0.43	0.79
	(0.23)	(0.14)	(0.20)	(0.40)
0.2	0.75	0.55	0.68	0.90
	(0.44)	(0.32)	(0.37)	(0.54)

(d) $\alpha = 0.1, \rho = 0.6$

μ_y \ σ_y	0.9	1.0	1.1	1.2
0	0.50	**0.10**	0.41	0.90
	(0.20)	**(0.08)**	(0.18)	(0.42)
0.1	0.71	0.40	0.63	0.94
	(0.36)	(0.21)	(0.30)	(0.52)
0.2	0.97	0.89	0.93	0.99
	(0.77)	(0.61)	(0.64)	(0.77)

Not surprisingly, both simulated powers increase with α and ρ. Indeed, the Kleijnen test proves to be greatly superior to the Stuart test. More detailed tables for the cases $\mu_y = 0$ and $\sigma_y = 1$ are presented in the Appendix.

The advantage of the Stuart test is its robustness: it is also applicable if the (bi)normality assumption is dropped. The Kleijnen test, on the other hand, uses the F-distribution and therefore heavily depends on normality. Therefore, in *part B* we investigate the behavior of the two tests when the normality assumption is not satisfied. Now, X and Y essentially have marginal gamma distributions; throughout part B, X and Y both have zero mean and unit variance.

For part B, two dependent variables with marginal gamma distributions were obtained as follows:

$$\left.\begin{array}{l} V \sim \Gamma(1,a) \\ W \sim \Gamma(1,b) \\ Z \sim \Gamma(1,c) \\ V, W, Z \text{ indep.} \end{array}\right\} \Rightarrow \left\{\begin{array}{l} V+W \sim \Gamma(1,a+b) \\ V+Z \sim \Gamma(1,a+c) \\ \rho(V+W, V+Z) = a/\sqrt{(a+b)(a+c)} \end{array}\right.$$

Next, the pair (X, Y) was obtained by standardization:

$$X = \frac{V+W-(a+b)}{\sqrt{a+b}}, \quad Y = \frac{V+Z-(a+c)}{\sqrt{a+c}} \tag{4.7}$$

As in part A, the variable X (and here Y as well) has mean 0 and variance 1, while $\rho = a/\sqrt{pq}$, where $p = a+b$, $q = a+c$. Note that for $p, q \to \infty$, the distributions of X and Y tend to the standard normal.

Table 2 shows the fractions (4.6); again, the sample size is 200 and the number of replications 2000. The two sub-tables relate to different levels of significance. Recall that the Kleijnen test is based on replacement of the original null hypothesis

X and Y are identically distributed

by (4.2):

X and Y have equal means and variances.

The two hypotheses are equivalent if and only if (X, Y) has a binormal distribution. In part B, the latter hypothesis is correct, whereas the former is correct only for $b = c$ (and $p = q$). Hence, for $b = c$, $\hat{\beta}$ should approximately equal α, whereas for $b \neq c$, higher values of $\hat{\beta}$ are to be preferred. As in Table 1, numbers between parentheses refer to the Stuart test and bold-faced number give the simulated size. More detailed results can be found in the Appendix.

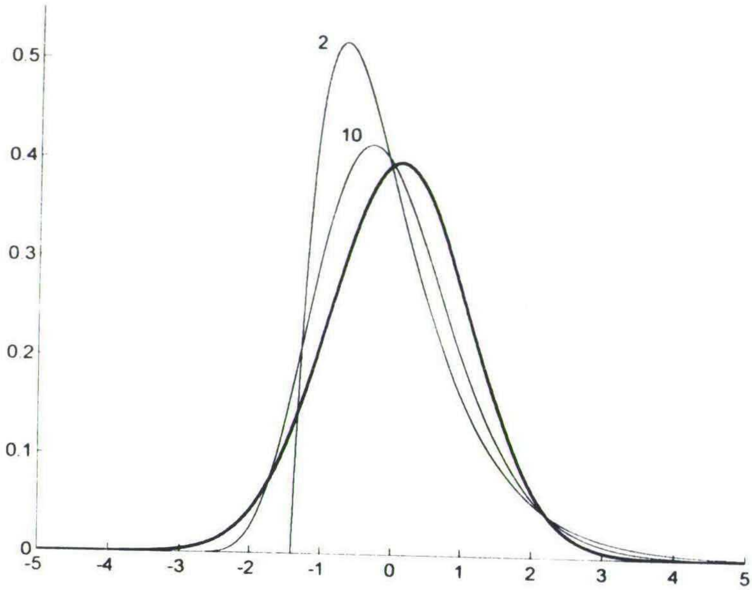
Table 2. Simulated powers $\hat{\beta}_K$ and $(\hat{\beta}_S)$; non-normal case.

(a) $\alpha = 0.05$ (b) $\alpha = 0.1$

		(a) ρ = 0	0.2	0.4			(b) ρ = 0	0.2	0.4
p	q				p	q			
2	2	**0.18**	**0.16**	**0.13**	2	2	**0.24**	**0.23**	**0.21**
		(0.04)	**(0.04)**	**(0.04)**			**(0.09)**	**(0.09)**	**(0.09)**
	6	0.12	0.12	0.11		6	0.20	0.18	0.19
		(0.55)	(0.53)	(0.52)			(0.69)	(0.67)	(0.67)
	10	0.13	0.12	0.11		10	0.20	0.17	0.17
		(0.71)	(0.72)	(0.67)			(0.81)	(0.82)	(0.80)
5	5	**0.12**	**0.11**	**0.08**	5	5	**0.18**	**0.17**	**0.14**
		(0.04)	**(0.05)**	**(0.04)**			**(0.10)**	**(0.10)**	**(0.09)**
	15	0.08	0.08	0.08		15	0.14	0.14	0.13
		(0.09)	(0.10)	(0.09)			(0.16)	(0.18)	(0.16)
	25	0.08	0.09	0.07		25	0.15	0.14	0.13
		(0.12)	(0.11)	(0.11)			(0.20)	(0.20)	(0.21)

The size of the Stuart test closely approximates the level of significance α; for the Kleijnen test, however, the size exceeds α with about 50% ($p = q = 5, \alpha = 0.1$) up to 300% ($p = q = 2, \alpha = 0.05$). As to the power, the Stuart test performs satisfactorily - although the power is low for $p = 5$. However, for the Kleijnen test the probability of rejecting H_0 even *decreases* for increasing q! The conclusion must be that the Kleijnen test should not be used whenever the binormality assumption is not (exactly) met. To show the deviations of normality considered here, Figure 3 presents two standardized gamma distributions $\Gamma(1, a)$, as well as the standard normal.

Figure 3. Standardized distributions $\Gamma(1,2)$ and $\Gamma(1,10)$ and standard normal distribution.



Our simulations were done by means of MATLAB. The procedure gamrnd has been used to generate random samples from the specified gamma distributions. The first two parameters of gamrnd may be matrices or scalars. We found that the version with matrices did not perform satisfactorily: sometimes, the variance in the simulated samples exceeded the true variance by 50%. Hence, we suspect this procedure of containing a bug, and advise the use of the version with scalars.

5 Discussion and conclusions

Simulation experiments can be subdivided in two categories, namely

- trace-driven, and
- independent

simulations. Experiments in the first category may have different objectives; these objectives lead to three different questions concerning the outcome X of the real life process and the outcome Y of the simulated process. In statistical terms, these three questions are:

A) is there a close agreement between individual values of X and Y?

B) is the joint distribution of (X, Y) symmetric?

C) are the (marginal) distributions of X and Y identical?

For independent simulation, only the last question is relevant.

For the real life mimicking experiment of case A), the correlation between X and Y should be high. We argued that the recently proposed Kleijnen test (KLEIJNEN et al., 1997) is therefore not suitable for this case. Instead, we introduced the validation measure M in (2.2) for real life mimicking simulation; this measure takes into account all three differences $\mu_y - \mu_x, \sigma_y - \sigma_x$ and $1 - \rho$.

For the two other questions, χ^2-statistics were proposed; for case B), this corresponds to a well-known test for symmetry.

For case C), the Kleijnen test statistic is a serious candidate: it should be applied whenever the distribution of (X, Y) is binormal. However, when binormality does not occur, the size of the Kleijnen test may greatly exceed the prescribed level of significance. This even holds for not too large deviations from binormality. Furthermore, the power of the Kleijnen test may even be lower than the size. The robustness of the Stuart test guarantees its satisfactory behaviour, regardless of binormality.

Acknowledgement

We thank Jack P.C. Kleijnen (Tilburg University) for his thorough and detailed comments.

Appendix

Simulated powers $\hat{\beta}_K$ and $\hat{\beta}_S$; normal case, $\sigma_y = 1$.

	$\hat{\beta}_K$					$\hat{\beta}_S$			
ρ α	0.0100	0.0250	0.0500	0.1000	μ_y	0.0100	0.0250	0.0500	0.1000
-0.4	0.0105	0.0240	0.0455	0.0920	0.0	0.0035	0.0200	0.0410	0.0925
	0.0220	0.0515	0.0985	0.1680	0.1	0.0100	0.0335	0.0645	0.1285
	0.1170	0.2015	0.2970	0.4185	0.2	0.0335	0.0805	0.1430	0.2510
	0.3540	0.4915	0.6040	0.7220	0.3	0.1230	0.2165	0.3210	0.4635
	0.6700	0.7835	0.8495	0.9070	0.4	0.3280	0.4665	0.5885	0.7040
	0.8900	0.9430	0.9670	0.9845	0.5	0.5885	0.7175	0.8120	0.8840
	0.9790	0.9900	0.9965	0.9995	0.6	0.8250	0.8965	0.9440	0.9730
0	0.0145	0.0315	0.0560	0.1085	0.0	0.0085	0.0265	0.0600	0.1080
	0.0465	0.0870	0.1340	0.2205	0.1	0.0240	0.0455	0.0850	0.1595
	0.1985	0.2985	0.4020	0.5285	0.2	0.0730	0.1355	0.2055	0.3145
	0.5285	0.6640	0.7550	0.8490	0.3	0.2275	0.3320	0.4510	0.5745
	0.8530	0.9165	0.9505	0.9765	0.4	0.5070	0.6435	0.7505	0.8465
	0.9780	0.9910	0.9965	0.9980	0.5	0.8070	0.8935	0.9400	0.9700
	0.9980	0.9995	0.9995	1.0000	0.6	0.9605	0.9795	0.9925	0.9965
0.2	0.0090	0.0210	0.0435	0.0970	0.0	0.0070	0.0145	0.0395	0.0925
	0.0510	0.0920	0.1595	0.2475	0.1	0.0160	0.0345	0.0810	0.1590
	0.2630	0.3945	0.5065	0.6340	0.2	0.0875	0.1615	0.2515	0.3780
	0.6845	0.7865	0.8620	0.9215	0.3	0.3150	0.4640	0.5815	0.7060
	0.9450	0.9730	0.9845	0.9935	0.4	0.6645	0.7925	0.8660	0.9285
	0.9955	0.9985	0.9990	1.0000	0.5	0.9180	0.9520	0.9775	0.9905
	1.0000	1.0000	1.0000	1.0000	0.6	0.9895	0.9955	0.9990	1.0000
0.4	0.0125	0.0335	0.0605	0.1095	0.0	0.0075	0.0225	0.0480	0.1005
	0.0740	0.1240	0.1940	0.2870	0.1	0.0225	0.0590	0.1040	0.1800
	0.3725	0.5115	0.6250	0.7385	0.2	0.1235	0.2120	0.3185	0.4510
	0.8285	0.9020	0.9420	0.9685	0.3	0.4520	0.5885	0.6985	0.8140
	0.9840	0.9955	0.9965	0.9985	0.4	0.8280	0.9050	0.9470	0.9750
	0.9985	0.9995	1.0000	1.0000	0.5	0.9790	0.9935	0.9955	0.9985
	1.0000	1.0000	1.0000	1.0000	0.6	0.9995	1.0000	1.0000	1.0000
0.6	0.0135	0.0275	0.0510	0.1015	0.0	0.0070	0.0255	0.0530	0.1080
	0.1080	0.1880	0.2710	0.3855	0.1	0.0365	0.0840	0.1390	0.2305
	0.6030	0.7265	0.8060	0.8810	0.2	0.2430	0.3700	0.4845	0.6355
	0.9625	0.9830	0.9925	0.9960	0.3	0.7115	0.8110	0.8860	0.9425
	1.0000	1.0000	1.0000	1.0000	0.4	0.9665	0.9880	0.9940	0.9980
	1.0000	1.0000	1.0000	1.0000	0.5	0.9985	1.0000	1.0000	1.0000
	1.0000	1.0000	1.0000	1.0000	0.6	1.0000	1.0000	1.0000	1.0000

Simulated powers $\hat{\beta}_K$ and $\hat{\beta}_S$; normal case, $\mu_y = 0$.

	$\hat{\beta}_K$					$\hat{\beta}_S$			
ρ α	0.0100	0.0250	0.0500	0.1000	σ_y	0.0100	0.0250	0.0500	0.1000
-0.4	0.9940	0.9980	1.0000	1.0000	0.7	0.7960	0.8855	0.9375	0.9630
	0.7215	0.8190	0.8755	0.9315	0.8	0.2175	0.3475	0.4790	0.6290
	0.1100	0.1990	0.2930	0.4110	0.9	0.0245	0.0560	0.1020	0.1855
	0.0080	0.0245	0.0485	0.0935	1.0	0.0085	0.0245	0.0505	0.0955
	0.0870	0.1535	0.2300	0.3440	1.1	0.0205	0.0490	0.0915	0.1710
	0.4715	0.5905	0.6905	0.7985	1.2	0.0890	0.1670	0.2580	0.3835
	0.8695	0.9320	0.9635	0.9845	1.3	0.2745	0.3910	0.5075	0.6425
	0.9935	0.9985	0.9990	1.0000	1.4	0.5205	0.6500	0.7520	0.8465
0	0.9845	0.9940	0.9960	0.9980	0.7	0.6815	0.7980	0.8710	0.9385
	0.6195	0.7375	0.8170	0.8890	0.8	0.1660	0.2790	0.3925	0.5325
	0.0880	0.1535	0.2395	0.3510	0.9	0.0260	0.0535	0.0955	0.1760
	0.0085	0.0260	0.0460	0.0970	1.0	0.0065	0.0195	0.0440	0.0985
	0.0720	0.1325	0.2000	0.2990	1.1	0.0175	0.0360	0.0850	0.1575
	0.3665	0.5060	0.6330	0.7560	1.2	0.0740	0.1370	0.2190	0.3430
	0.8055	0.8810	0.9285	0.9590	1.3	0.2190	0.3370	0.4585	0.5980
	0.9685	0.9850	0.9950	0.9980	1.4	0.4655	0.5960	0.7010	0.8080
0.2	0.9810	0.9925	0.9945	0.9980	0.7	0.7175	0.8180	0.8830	0.9305
	0.6220	0.7440	0.8245	0.8840	0.8	0.1795	0.2845	0.3890	0.5485
	0.1000	0.1785	0.2615	0.3735	0.9	0.0275	0.0620	0.1010	0.1685
	0.0090	0.0275	0.0530	0.0960	1.0	0.0070	0.0225	0.0465	0.0905
	0.0760	0.1325	0.2155	0.3165	1.1	0.0165	0.0355	0.0745	0.1535
	0.4085	0.5265	0.6350	0.7460	1.2	0.0700	0.1400	0.2230	0.3405
	0.8015	0.8760	0.9355	0.9640	1.3	0.2225	0.3505	0.4695	0.6060
	0.9725	0.9895	0.9940	0.9975	1.4	0.4615	0.6105	0.7215	0.8340
0.4	0.9930	0.9955	0.9990	0.9995	0.7	0.7620	0.8665	0.9275	0.9630
	0.7050	0.8015	0.8695	0.9250	0.8	0.1970	0.3235	0.4410	0.5785
	0.1155	0.1925	0.2805	0.3960	0.9	0.0215	0.0460	0.1020	0.1805
	0.0080	0.0245	0.0480	0.0980	1.0	0.0075	0.0185	0.0410	0.0895
	0.0910	0.1535	0.2310	0.3435	1.1	0.0245	0.0480	0.0890	0.1590
	0.4565	0.5990	0.7125	0.8150	1.2	0.0970	0.1670	0.2545	0.3925
	0.8730	0.9290	0.9590	0.9770	1.3	0.2730	0.3945	0.5080	0.6490
	0.9865	0.9970	0.9980	0.9990	1.4	0.5205	0.6575	0.7560	0.8510
0.6	1.0000	1.0000	1.0000	1.0000	0.7	0.8885	0.9420	0.9730	0.9850
	0.8380	0.9055	0.9450	0.9740	0.8	0.2900	0.4245	0.5475	0.6770
	0.1660	0.2485	0.3390	0.4730	0.9	0.0320	0.0750	0.1275	0.2280
	0.0080	0.0225	0.0460	0.0980	1.0	0.0085	0.0230	0.0480	0.1015
	0.1390	0.2220	0.3035	0.4295	1.1	0.0245	0.0600	0.1085	0.1880
	0.6380	0.7480	0.8405	0.8960	1.2	0.1205	0.2100	0.3110	0.4445
	0.9545	0.9770	0.9855	0.9945	1.3	0.3665	0.5150	0.6280	0.7295
	0.9995	1.0000	1.0000	1.0000	1.4	0.6515	0.7650	0.8410	0.9105

Simulated powers $\hat{\beta}_K$ and $\hat{\beta}_S$; non-normal case.

	$\hat{\beta}_K$						$\hat{\beta}_S$			
ρ α	0.0100	0.0250	0.0500	0.1000	p	q	0.0100	0.0250	0.0500	0.1000
0	0.0905	0.1325	0.1760	0.2420	2	2	0.0105	0.0250	0.0445	0.0870
	0.0660	0.1095	0.1535	0.2220		4	0.1110	0.1935	0.3055	0.4560
	0.0485	0.0820	0.1195	0.2025		6	0.2680	0.4180	0.5505	0.6860
	0.0480	0.0775	0.1220	0.1905		8	0.3665	0.5125	0.6425	0.7715
	0.0560	0.0800	0.1265	0.2010		10	0.4415	0.5880	0.7120	0.8130
0.2	0.0815	0.1145	0.1635	0.2300		2	0.0075	0.0180	0.0375	0.0890
	0.0560	0.0915	0.1365	0.2035		4	0.1015	0.2035	0.3190	0.4600
	0.0525	0.0840	0.1210	0.1815		6	0.2705	0.4045	0.5250	0.6695
	0.0565	0.0920	0.1290	0.2005		8	0.3470	0.4995	0.6275	0.7650
	0.0505	0.0805	0.1195	0.1720		10	0.4265	0.5860	0.7160	0.8225
0.4	0.0600	0.0960	0.1335	0.2065		2	0.0085	0.0185	0.0420	0.0920
	0.0480	0.0845	0.1250	0.1845		4	0.1145	0.2115	0.3150	0.4645
	0.0465	0.0790	0.1130	0.1870		6	0.2305	0.3790	0.5160	0.6710
	0.0480	0.0720	0.1070	0.1785		8	0.3580	0.5020	0.6325	0.7745
	0.0465	0.0750	0.1085	0.1700		10	0.4000	0.5505	0.6705	0.8005
0	0.0410	0.0790	0.1160	0.1840	5	5	0.0070	0.0170	0.0425	0.0980
	0.0340	0.0635	0.0950	0.1545		10	0.0120	0.0340	0.0650	0.1370
	0.0225	0.0485	0.0750	0.1400		15	0.0240	0.0525	0.0900	0.1630
	0.0270	0.0515	0.0865	0.1510		20	0.0205	0.0460	0.0990	0.1720
	0.0215	0.0485	0.0820	0.1485		25	0.0355	0.0695	0.1230	0.1975
0.2	0.0400	0.0675	0.1070	0.1675		5	0.0090	0.0245	0.0465	0.0990
	0.0310	0.0500	0.0790	0.1445		10	0.0150	0.0410	0.0690	0.1305
	0.0275	0.0485	0.0835	0.1380		15	0.0210	0.0505	0.1005	0.1750
	0.0220	0.0460	0.0735	0.1335		20	0.0230	0.0545	0.1010	0.1735
	0.0185	0.0455	0.0850	0.1410		25	0.0280	0.0680	0.1135	0.1990
0.4	0.0210	0.0480	0.0785	0.1370		5	0.0070	0.0195	0.0395	0.0860
	0.0260	0.0455	0.0810	0.1465		10	0.0155	0.0370	0.0775	0.1415
	0.0245	0.0465	0.0800	0.1295		15	0.0205	0.0525	0.0940	0.1645
	0.0185	0.0385	0.0725	0.1320		20	0.0210	0.0450	0.0895	0.1735
	0.0235	0.0395	0.0720	0.1320		25	0.0355	0.0640	0.1130	0.2125
0	0.0220	0.0475	0.0835	0.1390	10	10	0.0110	0.0250	0.0470	0.0940
	0.0155	0.0365	0.0665	0.1250		20	0.0115	0.0285	0.0535	0.1120
	0.0200	0.0435	0.0645	0.1230		30	0.0110	0.0260	0.0555	0.1175
0.2	0.0195	0.0400	0.0720	0.1315		10	0.0090	0.0250	0.0510	0.1000
	0.0270	0.0475	0.0730	0.1250		20	0.0085	0.0280	0.0625	0.1230
	0.0155	0.0365	0.0665	0.1255		30	0.0125	0.0255	0.0635	0.1140
0.4	0.0180	0.0440	0.0775	0.1370		10	0.0055	0.0185	0.0460	0.0950
	0.0205	0.0395	0.0690	0.1285		20	0.0120	0.0265	0.0585	0.1060
	0.0165	0.0365	0.0615	0.1200		30	0.0150	0.0350	0.0635	0.1210

References

BALCI, O. (1994), Validation, verification, and testing techniques throughout the life cycle of a simulation study, *Annals of Operations Research* 53, 121-173.

D'AGOSTINO, R.B. & STEPHENS, M.A. (1986), editors, *Goodness-of-fit techniques*, Marcel Dekker, New York.

DE FINETTI, B. (1974), *Theory of probability*, Wiley, New York.

KLEIJNEN, J.P.C. (1995), Verification and validation of simulation models, *European Journal of Operational Research* 82, 145-162.

KLEIJNEN, J.P.C., B.W.M. BETTONVIL & W.J.H. VAN GROENENDAAL (1997), Validation of trace-driven simulation models: a novel regression test, accepted by *Management Science*.

LAW, A.M. & KELTON, W.D. (1991), *Simulation modeling and analysis*, 2nd edition, McGraw-Hill, New York.

LINDGREN, B.W. (1976), *Statistical theory*, MacMillan, New York.

McNEMAR, O. (1947), Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* 12, 153-157.

PEGDEN, C.P., R.E. SHANNON & R.P. SADOWSKI (1990), *Introduction to simulation using SIMAN*, McGraw-Hill, New York.

STUART, A. (1955), A test for homogeneity of the marginal distributions in a two-way classification, *Biometrika*, 42, 412-416.

IN 1996 REEDS VERSCHENEN

713 Jeroen Suijs en Peter Borm
Cooperative games with stochastic payoffs: deterministic equivalents
Communicated by Prof.dr. A.J.J. Talman

714 Herbert Hamers
Generalized Sequencing Games
Communicated by Prof.dr. S.H. Tijs

715 Ursula Glunk en Celeste P.M. Wilderom
Organizational Effectiveness = Corporate Performance?
Why and How Two Research Traditions Need to be Merged
Communicated by Prof.dr. S.W. Douma

716 R.B. Bapat en Stef Tijs
Incidence Matrix Games
Communicated by Prof.dr. A.J.J. Talman

717 J.J.A. Moors, R. Smeets en F.W.M. Boekema
Sampling with probabilities proportional to the variable of interest
Communicated by Prof.dr. B.B. van der Genugten

718 Harry Webers
The Location Model with Reservation Prices
Communicated by Prof.dr. A.J.J. Talman

719 Harry Webers
On the Existence of Unique Equilibria in Location Models
Communicated by Prof.dr. A.J.J. Talman

720 Henk Norde en Stef Tijs
Determinateness of Strategic Games with a Potential
Communicated by Prof.dr. A.J.J. Talman

721 Peter Borm and Ben van der Genugten
On a measure of skill for games with chance elements
Communicated by Prof.dr. E.E.C. van Damme

722 Drs. C.J.C. Ermans RA and Drs. G.W. Hop RA
Financial Disclosure: A Closer Look
Communicated by Prof.drs. G.G.M. Bak RA

723 Edwin R. van Dam & Edward Spence
Small regular graphs with four eigenvalues
Communicated by Dr.ir. W.H. Haemers

724 Paul Smit
Inexact Iterations for the Approximation of Eigenvalues and Eigenvectors
Communicated by Prof.dr. M.H.C. Paardekooper

725 Tammo H.A. Bijmolt, Michel Wedel
A Monte Carlo Evaluation of Maximum Likelihood Multidimensional Scaling Methods
Communicated by Prof.dr. R. Pieters

726 J.C. Engwerda
On the open-loop Nash equilibrium in LQ-games
Communicated by Prof.dr. J.M. Schumacher

727 Jacob C. Engwerda en Rudy C. Douven
A game-theoretic rationale for EMU
Communicated by Prof.dr. J.E.J. Plasmans

728 Willem H. Haemers
Disconnected vertex sets and equidistant code pairs
Communicated by Prof.dr. S. Tijs

729 Dr. J.Ch. Caanen
De toekomst van de reserve assurantie eigen risico
Communicated by Prof.dr. A.C. Rijkers

730 Laurence van Lent
The Economics of an Audit Firm: The Case of KPMG in the Netherlands
Communicated by Prof.drs. G.G.M. Bak RA

731 Richard P.M. Builtjens, Niels G. Noorderhaven
The Influence of National Culture on Strategic Decision Making:
A Case Study of the Philippines

732 S. Tijs, F. Patrone, G. Pieri, A. Torre
On consistent solutions for strategic games
Communicated by Prof.dr. A.J.J. Talman

733 A.J.W. van de Gevel
From Strategic Trade Policy to Strategic Alliances in the Global Semiconductor Industry
Communicated by Prof.dr. A.B.T.M. van Schaik

734 M. Voorneveld, H. Norde
A Characterization of Ordinal Potential Games
Communicated by Prof.dr. S. Tijs

735 J.C. de Vos, B.B. van der Genugten
Fitting a stochastic model for Golden-Ten
Communicated by Dr.ir. A.A.F. van de Ven

736 Gert Nieuwenhuis
Ergodicity Conditions and Cesàro Limit Results for Marked Point Processes
Communicated by Prof.dr. B.B. van der Genugten

737 P.J.F.G. Meulendijks, D.B.J. Schouten
Dynamiek, Analyse en Politiek
Communicated by Prof.dr. A.H.J.J. Kolnaar

738 Antoon van den Elzen
Constructive application of the linear tracing procedure to polymatrix games
Communicated by Prof.dr. A.J.J. Talman

739 René van den Brink
Skewness of the Wage Distribution in a Firm and the Substitutability of Labor Inputs
Communicated by Prof.dr. P.H.M. Ruys

740 Sharon Schalk
General Equilibrium Model with a Convex Cone as the Set of Commodity Bundles
Communicated by Prof.dr. A.J.J. Talman

741 Jacob C. Engwerda
The Infinite Horizon Open-Loop Nash LQ-Game
Communicated by Prof.dr. J. M. Schumacher

IN 1997 REEDS VERSCHENEN

742 L.W.G. Strijbosch, J.J.A. Moors en A.G. de Kok
On the interaction between forecasting and inventory control
Communicated by Dr. R.M.J. Heuts

743 J. Vörös, J. Kriens en L.W.G. Strijbosch
A Note on the Kinks at the Mean Variance Frontier
Communicated by Dr. P.M. Kort

744 Edwin R. van Dam
Three-class association schemes
Communicated by Dr.ir. W.H. Haemers

Katholieke Universiteit Brabant
PO Box 90153
5000 LE Tilburg
The Netherlands