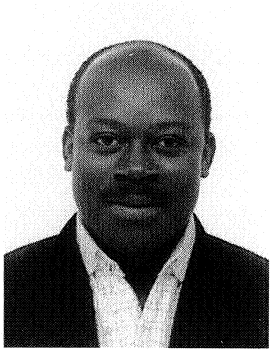
IN THIS ISSUE

A GENETIC MODEL FOR SICKLE CELL ANEMIA

In the first article of this issue, Goudjo Viho and Bernard Ycart propose a Markovian model to study the genetic drift of the disease Sickle Cell Anemia. This disease has had a very sad incidence on the own family of the first author, which motivated the present study.

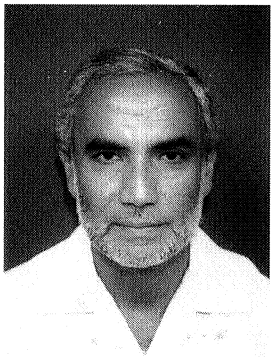


Goudjo Viho was born on 1968 in Lomé, Togo. After his undergraduate studies in Togo, he came to France and completed in 1996 a Ph.D. program in applied mathematics at the University Joseph Fourier of Grenoble. The subject of his thesis was the use of Markovian models in genetics.

AN OVERALL TEST FOR MULTIVARIATE NORMALITY

A multivariate normal distribution is often assumed in a multivariate problem of statistical inference. In the second article of this issue, C.R. Rao and Hydar Ali propose a method to verify if a p-variate sample of n observations is normally distributed. The technique consists in transforming the multivariate sample in a univariate sample of np observations and then in using the classical tests for univariate normality.

Hydar Ali was born on January 23, 1944 in California, Trinidad and Tobago. He obtained his Bachelor of Science in Mathematics at the University of the West Indies in 1965 and his Master of Science in Statistics at the University College London in 1966. In 1968, Hydar Ali received a M.Phil. in Applied University at the University of Sussex. He is currently working as a researcher at the University of the West Indies.



ON THE COMPLEMENTARITY OF FACTORIAL AND CLUSTER ANALYSIS

In the third article of this issue, Fabienne Benninghoff illustrates the complementarity of correspondence and cluster analysis with the spare-time activities of the Swiss residents.

Fabienne Benninghoff-Jeannerat was born on November 29, 1968 in Geneva, Switzerland. She obtained her Bachelor of Sciences in Sociology in 1992 from the University of Geneva and her postgraduate degree in Statistics in 1995 from the University of Neuchâtel. She is currently a chef of project with the Institute of Social and Preventive Medecine of Lausanne.



SUMS OF UNIFORMS, SUMS OF DISCRETE UNIFORMS AND THE CENTRAL LIMIT THEOREM

In a short and didactic paper Peter Thompson gives some illustrations on formulas involving sums of uniform random variables and the central limit theorem.

INGRAM OLKIN, STATISTICAL STATESMAN

In this issue, Ingram Olkin explains to the readers of *Student* the main reasons why matrix analysis is an important course to take in statistics. In the *Capturing History* section, homage is paid to this outstanding statistician.

FREQUENCY DISTRIBUTION OF ILLUSTRATED SCENES IN PERSIAN MANUSCRIPTS

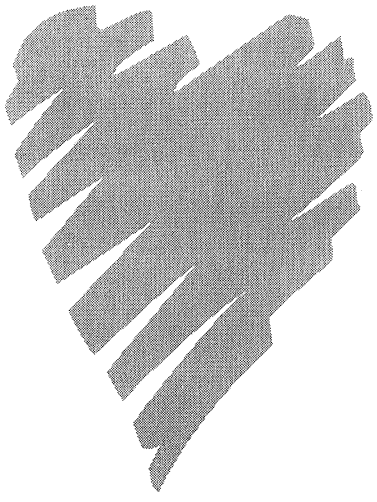
Dr. Farhad Mehran provides the readers of *Data & Statistics* with an application of statistics in Art in the field of Persian illustrated manuscripts. In his article, Dr. Mehran defines the concept of a random illustrated manuscript, and derives the probability distribution of scene frequencies in a large sample of manuscripts. He shows tha a particular distribution, coined the Diffused Waring distribution, closely fits the observed scene frequencies.

Acknowledgment of Referees' Services

The following individuals served as referees of papers for *Student* from March 1997 to October 1998. We are grateful for their outstanding and indispensable contribution.

The Editor

Andrew Barbour	Dennis Cook	David Cox
David Edwards	Jean-Pierre Gabriel	Marc Hallin
Jean-Marie Helbling	Norbert Henze	Marie Huskova
Jana Jureckova	Claudia Klüppelberg	Ludovic Lebart
Stephen Morgenthaler	Hans Nyquist	Ida Ruts
Stephen Senn	François Siegrist	Winfried Stute
Wenceslao Gonzalez-Manteiga		



Student (1998), Vol.2, No.4, 305-316

A Genetic Model for Sickle Cell Anemia

G. Viho and B. Ycart

LMC/IMAG, Grenoble, France

Abstract: The Sickle Cell Anemia is a genetic disease that affects populations of African origin around the world. It is encoded on a single locus gene with two alleles A and S. Only individuals with genotype SS develop the disease, whereas genotype AS induces a higher resistance to malaria. We propose a Markovian model to study the genetic drift of the disease, that takes into account the influence of Malaria as a selection factor. Using approximation results of Ethier and Nagylaki, we show that for large population sizes, the frequency of allele S can be described by a diffusion process the generator of which is given explicitly. An estimate of the expected time of disappearance of the disease is derived, and the influence of malaria as a selection factor is discussed.

Key words: Sickle Cell Anemia, Mathematical Genetics, Diffusion Approximation, Absorption Time.

1 Introduction

The Sickle Cell Anemia (SCA) (see [15] or [10] for general references) is a genetic disease caused by an abnormal type of hemoglobin, called hemoglobin S, and is inherited as an autosomal recessive trait. The disease occurs primarily in people of African heritage. Unfortunately, to this date SCA is beyond effective remedy, and kills a lot of young africans, although in the past ten years medical care has induced a certain improvement in the life expectancy of its victims. The production of hemoglobin S is a consequence of a single base mutation on the B-globin gene, and that gene can thus be considered to be one-locus, with two alleles, A (normal) and S (producing hemoglobin S). Only individuals who have inherited the allele S from both parents (genotype SS) will suffer from SCA. If the allele S is inherited from one parent only (genotype AS), the offspring will usually be without symp-

toms. Moreover, genotype AS yields immunity to malaria (see Luzzato [12]), which is another important cause of mortality in Africa.

Since SCA comes from a recessive trait, which is unlikely to be transmitted by parents of genotype SS (their life expectation being under 20 years), one may wonder why the disease did not disappear in time, by the effect of natural selection. It is generally believed that resistance to malaria of people with genotype AS, who thus have a higher probability to transmit their genes, may account for the permanence of SCA. The aim of this article is to propose a quantitative evaluation of the effect of malaria as a selection factor, on the incidence of SCA along successive generations.

Markov chain models for the evolution in time of gene frequencies in diploid populations have been proposed long ago, in particular by Moran [13] (see Ewens [9] for a general reference). The models are usually Markovian in the vector of (within-sex) genotypic frequencies. The importance of diffusion approximations in the treatment of these models for large populations has been recognized long ago, in particular by Kimura [11]. Contrarily to the basic monoecious Fisher-Wright model (Ewens [9] p. 19) the main mathematical difficulty in the treatment of these models is that the number of copies of an allele in the population does not have a Markovian evolution, being only a function of a higher dimension Markov process that describes the quantitative evolution of the genotypes among males and females. Diffusion approximations for general genetic models have been proposed in particular by Ethier and Nagylaki [7, 8]. Our goal is to apply their results to a model of evolution of the frequency of the allele S of SCA in a human population. The model is that of a single diallelic locus in a finite, panmictic, diploid dioecious population. It will be described precisely in Section 2. The application of Ethier and Nagylaki's [7, 8] results will be exposed in Section 3 and the generator of the limit diffusion will be computed. Expected times for the disappearance of the allele S are derived from the theoretical models in Section 4. Some elements of discussion for the obtained results will be given in the concluding Section 5.

2 The model

The model we use here is one of those introduced in the dioecious autosomal case by Moran [13]. It was then generalized by Ethier and Nagylaki [7, 8]. The generations are supposed to be non-overlapping, and for each generation, the numbers of males and females are fixed and equal to N_1 and N_2 respectively. The state of the system at the n-th generation is described by an integer valued vector with six coordinates:

$$s(n) \;=\; (m_0(n), m_1(n), m_2(n), f_0(n), f_1(n), f_2(n)) \;.$$

Here, for $i = 0, 1, 2$, $m_i(n)$ (respectively $f_i(n)$) denotes the number of males (resp. of females) at generation n having i copies of the allele S in their genotype (the possible genotypes are AA, AS, and SS). Obviously, for each generation n, one has

$$m_0(n) + m_1(n) + m_2(n) = N_1 \quad \text{and} \quad f_0(n) + f_1(n) + f_2(n) = N_2 \, .$$

The model is a homogeneous Markov chain on the state space

$$E = \{ s = (m_0, m_1, m_2, f_0, f_1, f_2) \in \mathbb{N}^6$$

$$\text{s.t. } m_0 + m_1 + m_2 = N_1 \, ; \; f_0 + f_1 + f_2 = N_2 \} \, .$$

For a state $s = (m_0, m_1, m_2, f_0, f_1, f_2)$, the proportions of allele S among males and females are denoted respectively by $x^{(m)}$ and $x^{(f)}$.

$$x^{(m)} = \frac{m_1 + 2m_2}{2N_1} \quad \text{and} \quad x^{(f)} = \frac{f_1 + 2f_2}{2N_2} \, . \tag{1}$$

In reproduction, male and female gametes are assumed to fuse at random (the so-called "multinomial" sampling scheme). At conception, the probability for an individual to have 0, 1, or 2 copies of allele S are denoted by $P_0^{(0)}$, $P_1^{(0)}$, and $P_2^{(0)}$ respectively. If the current proportions of the allele S among males and females are $x^{(m)}$ and $x^{(f)}$, the random mating hypothesis implies:

$$\begin{aligned} P_0^{(0)} &= (1 - x^{(m)})(1 - x^{(f)}) \, , \\ P_1^{(0)} &= x^{(m)}(1 - x^{(f)}) + (1 - x^{(m)})x^{(f)} \, , \\ P_2^{(0)} &= x^{(m)} x^{(f)} \, . \end{aligned} \tag{2}$$

However, due to selection, these probabilities have to be weighted by the so-called *viability factors* $w_i^{(m)}$ and $w_i^{(f)}$. For $i = 0, 1, 2$, $w_i^{(m)}$ (resp. $w_i^{(f)}$) measures the possibility for a male (resp. a female) with i copies of S in the genotype, to reach adult age. Let $P_i^{(m)}$ (resp. $P_i^{(f)}$) denote the probabilities for males (resp. females) to reach the mating age in the next generation with i copies of S. If the probability at conception is $P_i^{(0)}$, then the probabilities at mating age for males and females are:

$$\begin{aligned} P_i^{(m)} &= \frac{w_i^{(m)} P_i^{(0)}}{w_0^{(m)} P_0^{(0)} + w_1^{(m)} P_1^{(0)} + w_2^{(m)} P_2^{(0)}} \, , \\ P_i^{(f)} &= \frac{w_i^{(f)} P_i^{(0)}}{w_0^{(f)} P_0^{(0)} + w_1^{(f)} P_1^{(0)} + w_2^{(f)} P_2^{(0)}} \, . \end{aligned} \tag{3}$$

After regulation, the genotypes of the N_1 males and the N_2 females are assumed to be chosen independently with probabilities $P_i^{(m)}$ and $P_i^{(f)}$. In

other terms, the probability distributions of the vectors (m_0, m_1, m_2) and (f_0, f_1, f_2) are multinomial with respective parameters $(N_1, P_0^{(m)}, P_1^{(m)}, P_2^{(m)})$ and $(N_2, P_0^{(f)}, P_1^{(f)}, P_2^{(f)})$.

We can now explicitly describe the transition probabilities of our homogeneous Markov chain on E. If $s = (m_0, m_1, m_2, f_0, f_1, f_2)$ and $s' = (m'_0, m'_1, m'_2, f'_0, f'_1, f'_2)$ are two states in E, then the transition probability $p_{s,s'}$ from s to s' is:

$$\frac{N_1!}{m'_0!m'_1!m'_2!}\left(P_0^{(m)}\right)^{m'_0}\left(P_1^{(m)}\right)^{m'_1}\left(P_2^{(m)}\right)^{m'_2}\cdot$$

$$\frac{N_2!}{f'_0!f'_1!f'_2!}\left(P_0^{(f)}\right)^{f'_0}\left(P_1^{(f)}\right)^{f'_1}\left(P_2^{(f)}\right)^{f'_2}, \tag{4}$$

where for $i = 0, 1, 2$, the probabilities $P_i^{(m)} = P_i^{(m)}(s)$ and $P_i^{(f)} = P_i^{(f)}(s)$ are deduced from $s = (m_0, m_1, m_2, f_0, f_1, f_2)$ through formulas (1), (2) and (3).

The model that has just been described depends *a priori* on 8 parameters, namely the population sizes N_1, N_2, and the viability factors $w_i^{(m)}$, $w_i^{(m)}$. Since the expressions (3) are homogeneous in the viability factors for each sex, only 4 of these factors are independent. Two of them can be arbitrarily fixed to 1. For instance

$$w_0^{(m)} = w_0^{(f)} = 1\,.$$

The statistical estimation of viability factors with real data is not an easy task, and is frequently overlooked in the genetics literature (*cf.* [9]). We shall briefly discuss this problem in Section 4 for our example.

The model that has just been presented contains several grossly simplified hypotheses which do not hold for human populations, such as constant size population, non overlapping generations, random mating and so on. These hypotheses are common in genetic models, and it is generally believed that if the models do not reflect exactly the reality, they are still able to grasp at least the correct orders of magnitude in the quantitative evolution of genetic drifts.

3 Diffusion approximation

Our goal in this section is to establish the convergence, as the population size tends to infinity, of the proportion of the allele S to a diffusion process. For this, we shall denote by N the population size, and consider that the numbers of males and females are functions of N that tend to infinity with

N in such a way that the proportion between sexes is preserved.

$$N_1 = N_1(N) \longrightarrow \infty \quad ; \quad N_2 = N_2(N) \longrightarrow \infty \quad ; \quad \frac{N_1(N)}{N_2(N)} \longrightarrow \gamma \neq 0 \, .$$

It is customary in genetics to define the *effective population size* as the harmonic mean of N_1 and N_2:

$$N_e = 2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} .$$

If

$$S^N(n) = (m_0(n), m_1(n), m_2(n), f_0(n), f_1(n), f_2(n))$$

is the current state of the Markov chain at generation n, the proportions of the allele S among males, among females, and their average will be denoted by $X^{(m)}(n)$, $X^{(f)}(n)$, and $X^N(n)$.

$$X^{(m)}(n) = \frac{m_1(n) + 2m_2(n)}{2N_1} \quad ; \quad X^{(f)}(n) = \frac{f_1(n) + 2f_2(n)}{2N_2} \, ,$$

$$X^N(n) = \frac{1}{2}(X^{(m)}(n) + X^{(f)}(n)) \, .$$

Obviously $\{X^N(n) \, , \, n \geq 0\}$ is not a Markov chain. However, it is possible to show that up to a proper rescaling of time, it converges to a diffusion process. Indeed the main result of Ethier and Nagylaki [8], Theorem 2.1 p. 528 can be applied. That theorem is recalled below.

Theorem 1 *Let $\{S^N(n) \, ; \, n \geq 0\}$ be a time-homogeneous Markov chain in a metric space E_N.*

Suppose $\{X^N(n) \, ; \, n \geq 0\}$ and $\{Y^N(n) \, ; \, n \geq 0\}$ are two sequences of random variables such that:

(a) *$X^N(n) = \varphi(S^N(n))$; $Y^N(n) = \psi(S^N(n))$ for all $n \geq 0$ where φ and ψ are Borel mesurable functions.*

(b) *$\{(X^N(n), Y^N(n)) \, ; \, n \geq 0\}$ is a time-homogeneous Markov chain.*

Let (ε_N) and (δ_N) be two time scales such that

(a)

$$\lim_{N \to +\infty} \delta_N = \delta_\infty \geq 0,$$

(b)

$$\lim_{N\to+\infty}(\frac{\varepsilon_N}{\delta_N}) = 0,$$

Let $z \in E_N$, $x = \varphi(z)$ $y = \psi(z)$. Suppose that, as N tends to infinity, if the Markov chain S^N starts from $S^N(0) = z$, then the variances and means of $X^N(1)$ are asymptotically equivalent to $\varepsilon_N a(x,y)$ and $\varepsilon_N b(x,y)$ respectively, whereas those of $Y^N(1)$ are equivalent to 0 and $\delta_\infty c(x,y)$, uniformly in $z \in G_{K,N}$, where

$$G_{K,N} = \{\, z \in E_N \,;\ |x| \leq K\,,\ |y| \leq K \,\}\,.$$

($|.|$ denotes the Euclidean norm of the corresponding space).

Suppose further that $c(x,0) = 0$ for all $x \in \mathbb{R}$ and the zero solution of the differential equation (if $\delta_\infty = 0$)

$$\frac{d}{dt}f(t,x,y) = c(x, f(t,x,y))\,, \quad f(0,x,y) = y\,,$$

or difference equation (if $\delta_\infty > 0$)

$$\delta_\infty^{-1}[f(k+1,x,y) - f(k,x,y)] = c(x, f(k,x,y))\,, \quad f(0,x,y) = y\,,$$

is asymptotically stable, uniformly in (x,y) on compact sets.

Assume that for each $T > 0$

$$\lim_{K\to\infty}\liminf_{N\to\infty} Prob\left[\, S^N\left(\left[\frac{t}{\varepsilon_N}\right]\right) \in G_{K,N} \text{ for } 0 \leq t \leq T\right] = 1\,.$$

($[.]$ denotes the integer part).

Assume finally that a, b and c are continuous.

Then one has:

(1) $\{X^N\left(\left[\frac{t}{\varepsilon_N}\right]\right)\,;\, t \geq 0\}$ *converges weakly to a diffusion process $\{X_t\,;\, t \geq 0\}$ with diffusion and drift coefficients respectively equal to $a(x,0)$ and $b(x,0)$.*

(2) $\{Y^N\left(\left[\frac{t}{\delta_N}\right]\right)\,; t \geq 0\}$ *tends to zero in probability for every $t > 0$.*

Ethier and Nagylaki describe the application of their results to standard genetic models and in particular to the one described in Section 2

([7], p. 25). The main difficulty is to give appropriate definitions for the quantities $Y^N(n)$, ε_N, and δ_N of Theorem 1. These definitions are the following:

$$\begin{aligned}
\varepsilon_N &= (2N_e)^{-1}, \\
\delta_N &= 1, \\
Y^N(n) &= (Y_1^N(n), Y_2^N(n), Y_3^N(n), Y_4^N(n), Y_5^N(n), Y_6^N(n), Y_7^N(n)), \\
Y_1^N(n) &= \frac{X^{(m)}(n) - X^{(f)}(n)}{2}, \\
Y_2^N(n) &= \frac{m_0}{N_1} - (1 - X^N(n))^2, \\
Y_3^N(n) &= \frac{m_1}{N_1} - 2X^N(n)(1 - X^N(n)), \\
Y_4^N(n) &= \frac{m_2}{N_1} - (X^N(n))^2, \\
Y_5^N(n) &= \frac{f_0}{N_2} - (1 - X^N(n))^2, \\
Y_6^N(n) &= \frac{f_1}{N_2} - 2X^N(n)(1 - X^N(n)), \\
Y_7^N(n) &= \frac{f_2}{N_2} - (X^N(n))^2.
\end{aligned}$$

The infinitesimal variance and mean of $X^N(1)$ have been computed, and lead to the following expressions for the drift and diffusion coefficients of the limit diffusion process $\{X_t\,;\, t \geq 0\}$:

$$\begin{aligned}
a(x,0) &= x(1-x), \\
b(x,0) &= x(1-x)(\sigma_2 x + \sigma_1(1-2x) - \sigma_0(1-x)),
\end{aligned} \tag{5}$$

where for $i = 0, 1, 2$:

$$\begin{aligned}
\sigma_i &= \frac{\sigma_i^{(m)} + \sigma_i^{(f)}}{2}, \\
\sigma_i^{(m)} &= 2N_e(w_i^{(m)} - 1), \\
\sigma_i^{(f)} &= 2N_e(w_i^{(f)} - 1).
\end{aligned}$$

Thus for a large population size the frequency of allele S can be described by a diffusion process $\{X_t\,;\, t \geq 0\}$ with infinitesimal generator:

$$G = \frac{1}{2}a(x,0)\frac{\partial^2}{\partial x^2} + b(x,0)\frac{\partial}{\partial x}. \tag{6}$$

4 Time to absorption

In this section, we propose to evaluate the expected time to disappearance of the allele S, using the diffusion approximation of the previous section.

The Markov chain model described in Section 2 has obviously two absorbing states

$$(N_1, 0, 0, N_2, 0, 0) \quad \text{and} \quad (0, 0, N_1, 0, 0, N_2) \; .$$

The first one corresponds to the disappearance of the allele S, the second one to a complete invasion. Whatever the initial state, the Markov chain will eventually get trapped into one of these two states. However, if the initial proportion of S in the population is low, the probability of a complete invasion will be small. We shall neglect here this catastrophic possibility, and identify the time to disappearance to the absorption time of the chain.

Regarding the proportions $X^N(.)$ and their diffusion approximation, the absorption times are defined as the first hitting time of the absorbing set $\{0, 1\}$.

$$\begin{aligned} \tau^N &= \inf\{\, t \geq 0 \; s.t. \; X^N\left(\left[\frac{t}{\varepsilon_N}\right]\right) \in \{0,1\} \,\} \; , \\ \tau &= \inf\{\, t \geq 0 \; s.t. \; X(t) \in \{0,1\} \,\} \; . \end{aligned}$$

It is not obvious that the expectation of τ^N should converge to that of τ as N tends to infinity. This has been proved by Ethier and Kurtz [6] (Theorem 2.4 and Corollary 2.6 p. 419).

Theorem 2

$$\lim_{N \to \infty} I\!E[\tau^N \mid S^N(0) = z] \; = \; I\!E[\tau \mid X(0) = \varphi(z)] \; . \tag{7}$$

The computation of the expected absorption time of a diffusion process is easy, using the following classical result ([3]), recalled also in [6] (Proposition 2.8 p. 420).

Theorem 3 *The expected absorption time* $I\!E[\tau | X(0) = x]$ *is the unique* $\mathcal{C}[0,1] \cap \mathcal{C}^2(]0,1[)$ *solution of the differential equation:*

$$\frac{1}{2} a(x,0) \frac{d^2 f}{dx^2} + b(x,0) \frac{df}{dx} = -1 \; , \tag{8}$$

with boundary conditions:

$$I\!E[\tau | X(0) = 0] = I\!E[\tau | X(0) = 1] = 0 \; .$$

μ_0	μ_1	μ_2
0.012	0.005	0.041

Table 1: Death probabilities before maturity age in Lomé.

($a(x,0)$ and $b(x,0)$ are the diffusion and drift coefficients defined in Section 3; $\mathcal{C}[0,1]$ is the set of continuous functions on $[0,1]$; $\mathcal{C}^2(]0,1[)$ is the set of twice derivable functions on $]0,1[$).

This theorem is applied to the model described in Section 3 and the integral form of the solution of the differential equation (4.1) is:

$$I\!E[\tau|X(0)=x] = \int_0^x e^{-\lambda(q)} \int_q^{\frac{1}{2}} \frac{2e^{\lambda(r)}}{r(1-r)} dr dq$$

$$- \frac{\int_0^1 e^{-\lambda(q)} \int_q^{\frac{1}{2}} \frac{2e^{\lambda(r)}}{r(1-r)} dr dq}{\int_0^1 e^{-\lambda(q)} dq} \int_0^x e^{-\lambda(q)} dq. \tag{9}$$

where

$$\lambda(x) = \sigma_2 x^2 - \frac{\sigma_1}{2}(1-2x)^2 . \tag{10}$$

A similar expression is given by Ethier and Kurtz [6], p. 420. Thus, the mean time to disappearance of the S allele will be numerically computed after the determination of selection coefficients.

For an african town like Lomé (Togo), the population size is about $N = 802000$ with $N_1 = 387000$ men and $N_2 = 415000$ women (the data are taken from [1]). The effective population size is $N_e = 400512$. The proportion of allele S in the population is 0.161 [10]. We have to explain first how selection factors have been estimated. Let μ_i denote the proportion of people, having i copies of the allele S in their genotype, who die before maturity age. These quantities are approximately the same for men and women and are given in Table 1 (the data are taken from [2]).

The selection coefficient w_i is the relative likelihood for an individual having i allele S in his genotype to reach adult age (see Section 2). It is estimated by:

$$w_i = \frac{(1-\mu_i)}{(1-\mu_0)} .$$

The results are plugged into the integral expression (9) of the mean time to absorption, which can then be evaluated numerically, using composite Newton-Cotes formulas. One gets for instance:

$$E[\tau|X(0)=0.161] = 1.286 \times 10^{797}.$$

μ_0	$\mathbb{E}[\tau \mid X(0) = 0.161]$	Generations
0.012	1.286×10^{797}	1.030×10^{803}
0.011	5.688×10^{597}	4.556×10^{603}
0.010	2.686×10^{423}	2.151×10^{429}
0.009	8.131×10^{275}	6.513×10^{281}
0.008	1.473×10^{157}	1.180×10^{163}
0.007	2.540×10^{69}	2.035×10^{75}
0.006	1.451×10^{15}	1.162×10^{21}
0.005	1.154×10^{-2}	9244
0.004	5.440×10^{-3}	4358
0.003	3.494×10^{-3}	2799
0.002	2.625×10^{-3}	2103
0.001	2.122×10^{-3}	1700

Table 2: Expected times to disappearance of SCA.

In practice, this very large number means that SCA will never disappear if the present parameters are not modified.

Since genotype AS induces a higher resistance to malaria, one can assume that if malaria was eradicated, people would have the same chances to reach maturity with the genotype AS or AA. Progress in the struggle against malaria should decrease the probability for an individual with genotype AA to die before maturity age. If μ_0 decreases, the selection coefficients σ_1 and σ_2 decrease also. Table 2 shows the decrease of the expected time to absorption, as μ_0 decreases. The rightmost column gives the value of that disappearing time in number of generations, obtained after multiplication of the former by $2N_e = 801024$. Although still quite large, the disappearing time decreases very sharply, up to an apparent threshold lying around $\mu_0 = 0.005$, which is precisely the current value of μ_1 (Table 1).

5 Conclusion

We are very much aware of the weaknesses of this study, due to the simplified model we used and also to the difficulty in obtaining precise estimates for the viability coefficients. In spite of these weaknesses, we believe that strong evidence has been provided for two facts about the eradication of SCA.

Firstly, natural extinction seems hopeless as long as the selection effect of malaria will last. Secondly it is likely that any progress in the struggle against malaria, by increasing the proportion of genotype AA in the population, should result in a considerable decrease of the impact of SCA in the population.

Résumé: L'Anémie Falciforme, ou Drépanocytose est une maladie génétique qui affecte les populations d'origine africaine. Elle est codée par un gène à deux formes allèles, A et S. Seuls les individus dont le génotype est SS développent la maladie, tandis que le génotype AS induit une résistance accrue à la malaria. Nous proposons un modèle markovien pour étudier l'évolution de la Drépanocytose, prenant en compte l'influence de la malaria comme facteur de sélection. En utilisant des résultats d'approximation de Ethier et Nagylaki, nous montrons que pour une grande taille de population, l'évolution temporelle de la fréquence de la forme allèle S peut être décrite par un processus de diffusion, dont le générateur est donné explicitement. On en déduit une estimation du temps de disparition de la maladie, prenant en compte l'influence de la malaria comme facteur de sélection.

Mots clés: Anémie Falciforme, génétique mathématique, approximation de diffusion, temps d'absorbtion

References

[1] *Annales de la bibliothèque de l'Unité de Recherche Démographique (U.R.D.) du Togo* (1994). Lomé, Togo.

[2] *Annuaires des statistiques sanitaires du Togo* (1993). (C.H.U) Lomé, Togo.

[3] Barucha-Reid, A.T. (1960). *Elements of the theory of Markov processes and their Applications.* McGraw-Hill, London.

[4] Encyclopædia Universalis (1988). *Les chiffres du monde.* Paris.

[5] Danis, M. and Mouchet, J. (1991). *Paludisme. Universités Francophones.* ELLIPSES.

[6] Ethier, S.N. and Kurtz, T.G. (1986). *Markov processes: characterization and convergence.* Wiley, New York.

[7] Ethier, S.N. and Nagylaki, T. (1980). Diffusion approximation of Markov chains with two time scales and applications to population genetics. *Adv. Appl. Probab.* 12, 14–49.

[8] Ethier, S.N. and Nagylaki, T. (1988). Diffusion approximation of Markov chains with two time scales and applications to population genetics, II. *Adv. Appl. Probab.* 20, 525–545.

[9] Ewens, W.J. (1979). *Mathematical population genetics.* Biomathematics 9, Springer-Verlag, Berlin.

[10] Galactéros, F.G. and Dormont, S. (1990). *Drépanocytose et santé publique.* CIE–INSERM, Paris.

[11] Kimura, M. (1964). Diffusion models in population genetics. *J. Appl. Probab.* 1, 177–232.

[12] Luzzato, L. (1979). Genetic of red cells and susceptibility to malaria. *Blood* 54, 961–976.

[13] Moran, P.A.P. (1958). Random processes in genetics. *Proc. Camb. Phil. Soc.* 54, 60–71.

[14] Nagylaki, T. (1977). *Selection in One- and Two-Locus systems.* L.N. in Biomathematics 15, Springer-Verlag, Berlin.

[15] Serjeant, S. and Graham, R (1992). *Sickle Cell Disease.* Oxford University Press.

[16] Viho, G. (1996). *Etude de modèles markoviens en génétique et calcul des temps d'absorption.* Thèse, Université J. Fourier, Grenoble.

[17] World Health Organization (1977). Manual of international statistical classification of diseases, injuries and causes of death. Ninth Revision. Geneva.

Student (1998), Vol.2, No.4, 317-324

An Overall Test For Multivariate Normality

C.R. Rao

Pennsylvania State University, University Park, USA

Hydar Ali

The University of West Indies, St. Augustine, Trinidad & Tobago

Abstract: There are a number of methods in the statistical literature for testing whether observed data come from a multivariate normal (MVN) distribution with an unknown mean vector and covariance matrix. Let $X_1, \ldots, X_n$ be an iid sample of size n from a p-variate normal distribution. Denote the sample mean and sample variance-covariance matrix by $\bar{X}$ and S respectively. Most of the tests of multivariate normality are based on the result that $Y_i = S^{-1/2}(X_i - \bar{X}), i = 1, \ldots, n$, are asymptotically iid as p-variate normal with zero mean vector and identity covariance matrix. Tests developed by Andrews et al., Mardia and others are direct functions of Y_i's. We note that the np components of all the Y_i's put together can be considered as an asymptotically iid sample of size $N = np$ from a univariate normal (UVN) distribution with zero mean and unit variance. We test this hypothesis using any well known test based on N independent observations for univariate normality. In particular we suggest univariate skewness and kurtosis tests, which are sensitive to deviations from normality, and also use of normal probability plots.

Key words and phrases: Skewness and kurtosis tests, Univariate normality tests, Multivariate normality tests, Normal probability plot.

1 Introduction

The assumption of multivariate normality (MVN) underlies many important techniques in multivariate analysis. When the sample size is large one can use test criteria based on (or optimal for) MVN but refer them to their

asymptotic distributions which are independent of the underlying distribution of the observations. Thus, if one computes Hotelling's T^2 to test the equality of means of p variables in two populations, a constant times T^2 can be considered as a chisquare on p degrees of freedom although the measurements are not MVN, provided the sample sizes for the two populations are large. In small samples such a procedure may give misleading results. However, simulation results show that the p-values based on the assumption of MVN may not be widely off if the underlying distribution is close to MVN. We suggest that in practice it is advisable to carry out a test for MVN on observed data before applying any methodology based on MVN. If non-normality is detected we may wish to see if a suitable transformation can restore multivariate normality or explore the possibility of using nonparametric procedures.

There are numerous tests available in the statistical literature for testing MVN. For general surveys of these tests, the reader is referred to papers by Mardia (1980), Andrews, Gnanadesikan and Warner (1971), Cox and Small (1978), Baringhaus and Henze (1988), Henze and Wagner (1997) and books by Jobson (1992) and others.

No single test is more powerful than the others to detect all kinds of departures from multivariate normality. However, appropriate tests may be constructed to detect specific departures from normality, an example of which can be found in Cox and Small (1978). The tests suggested by Andrews et al., Mardia and others are based on particular properties of the MVN distribution.

In this paper, we suggest an overall test which can be routinely implemented in the initial data analysis (or cross examination of data as discussed in Rao, 1997) for choosing a model for inferential data analysis on mean values, construction of discriminant functions and related problems. The proposed test[1] appears to be more sensitive than some others in the data sets we have examined. However, as in the case of other tests, its efficiency depends on particular types of departure from MVN.

2 The proposed test

Let $X_1, \cdots, X_n$ be an iid sample of size n from a p-variate population and denote the $p \times n$ matrix $(X_1, \cdots, X_n)$ by X. Then the sample mean and estimated variance-covariance matrix are

$$\bar{X} = n^{-1}X1, \quad S = (n-1)^{-1}X(I - n^{-1}11')X' \tag{1}$$

where 1 is the n-vector of unites.

[1]Editor's Note: The test statistic proposed in this article is not affine invariant.

Consider the spectral decomposition of S

$$S = \lambda_1^2 P_1 P_1' + \cdots + \lambda_p^2 P_p P_p' = P\Lambda^2 P' \tag{2}$$

where the λ_i^2's are the eigen values and the P_i's are the eigen vectors, $P = (P_1, \cdots, P_p)$ and Λ^2 is a diagonal matrix with λ_i^2 as its i-th diagonal element. Then

$$S^{-1/2} = \lambda_1^{-1} P_1 P_1' + \cdots + \lambda_p^{-1} P_p P_p' = P\Lambda^{-1} P' \tag{3}$$

is a positive definite square root of S^{-1}. Further let

$$Y_i = S^{-1/2}(X_i - \bar{X}), \qquad i = 1, \cdots, n. \tag{4}$$

Then under the assumption that the X_i's are iid variables from $N_p(\mu, \sum)$, i.e., a p-variate normal distribution with mean μ and covariance matrix $\sum$, smooth functions of moments of Y_i, tend to have the same distribution as the corresponding functions of $Z_i = \sum^{-1/2}(X_i - \mu)$ as $n \to \infty$. We consider the np-vector

$$Y' = (Y_1', \cdots, Y_n') \tag{5}$$

as approximately distributed as $N_{np}(0, I)$, i.e., treat all the $N = np$ components of Y as iid variables from $N_1(0, 1)$, and apply a test for univariate normality based on N observations. Thus assessing MVN is reduced to a test for UVN (univariate normality). We believe such a test will be more sensitive as it is based on a large number of observations. Other authors cited in the introduction use different functions of Y_i in constructing tests of MVN. Andrews et al. (1971) use $Y_i' Y_i, i = 1, \cdots, n$, as independent chisquares on p degrees of freedom. Mardia (1980) uses $Y_i' Y_j$ and $Y_i' Y_i$ in defining multivariate skewness and kurotsis respectively and suggests tests based on them. The test proposed by Baringhaus and Henz (1988) is a function of $\|Y_i\|^2$ and $\|Y_i - Y_j\|^2$. These tests are invariant for a linear transformation of the X_i, unlike the test we propose. But it is believed that our tests may be more sensitive.

In this paper we consider tests of UVN based on $\sqrt{b_1}$ and b_2, the measures of skewness and kurtosis computed from the N observations in (5). One could also use the Shapiro-Wilk W test by considering the normal approximation

$$z = [(1 - W)^\lambda - \mu]/\sigma \tag{6}$$

where λ, μ and σ can be obtained for given n from Table 2 of Royston (1982, p. 119) or the normal approximation by using the Johnson (1949) S_b family

$$z = \hat{\gamma} + \hat{\delta} \log[(W - \hat{\epsilon})/(1 - W)]$$

where $\hat{\lambda}, \hat{\delta}$ and $\hat{\epsilon}$ are obtained from Table 1 of Shapiro and Wilk (1968) if $n \leq 50$. If $n > 50$, these estimates can be obtained using the results of Shapiro and Francia (1972) and Royston (1982). [See Srivastava and Hui (1987, p. 16)]. Other possible tests for univariate normality are those proposed by Epps and Pulley (1983) and Puri and Rao (1975).

The tests based on skewness and kurtosis are described in D'Agostino et al. (1990). The computational steps are reproduced below. Denote the $N = np$ observations in (5) by $U_1, \cdots, U_N$ and by m_r, the r-th corrected moment

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(U_i - \bar{U})^r$$

where $\bar{U}$ is the average of $U_1, \cdots, U_N$.

For skewness test, compute, $\sqrt{b_1} = m_3/m_2^{3/2}$,

$$\begin{aligned}
Y &= \sqrt{b_1}\left[\frac{(N+1)(N+3)}{6(N-2)}\right]^{1/2}, \\
\beta_2(\sqrt{b_1}) &= \frac{3(N^2+27N-70)(N+1)(N+3)}{(N-2)(N+5)(N+7)(N+9)}, \\
W^2 &= -1 + [2(\beta_2(\sqrt{b_1})-1)]^{1/2}, \quad \delta = (\log W)^{-1/2}, \\
\alpha &= \left\{2/(W^2-1)\right\}^{1/2}, \\
Z(\sqrt{b_1}) &= \delta \log\left\{Y/\alpha + [(Y/\alpha)^2+1]^{1/2}\right\} \sim N(0,1),
\end{aligned} \tag{7}$$

where the logarithm is to the base e. The test statistic for assessing skewness is (7), which can be referred to a normal table for obtaining the p value.

For kurtosis test compute, $b_2 = m_4/m_2^2$

$$\begin{aligned}
E(b_2) &= \frac{3(N-1)}{N+1}, \; V(b_2) = \frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}, \\
\sqrt{\beta_1(b_2)} &= \frac{6(N^2-5N+2)}{(N+7)(N+9)}\left[\frac{6(N+3)(N+5)}{N(N-2)(N-3)}\right]^{1/2} \\
x &= \frac{b_2 - E(b_2)}{\sqrt{V(b_2)}}, \; A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}}\left[\frac{2}{\sqrt{\beta_1(b_2)}} + (1+\frac{4}{\beta_1(b_2)})^{1/2}\right], \\
Z(b_2) &= [(1-\frac{2}{9A}) - (\frac{1-2/A}{1+x\sqrt{2/(A-4)}})^{1/3}]/\sqrt{2/9A}.
\end{aligned} \tag{8}$$

Test for kurtosis can be carried out by referring $Z(b_2)$ to a normal table with mean zero and variance unity. The normal distributions in (7) and (8) are valid only asymptotically and the tests are recommended only when

large samples are available. The expressions for $E(b_2), V(b_2), \beta_2(\sqrt{b_1})$ and $\beta_1(b_2)$ are exact if $U_1, \cdots, U_N$ are iid from $N_1(\mu, \sigma^2)$. They are approximate in the present case since $U_1, \cdots, U_N$ are iid only asymptotically.

A combined test for skewness and kurtosis is to use

$$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2) \tag{9}$$

as chisquare on two degrees of freedom. The tests based on the N observations (5) are denoted by T_1 in the Tables 1, 2 and 3 in Section 3 of the paper.

The asymptotic theory of the tests (7), (8) and (9) is the same as that of Mardia's test for skewness and kurtosis. The relative merits of our tests have to be examined.

3 Other transformations to independent variables

It is well known that if μ and $\sum$ are known, the p-vector variable $X \sim N_p(\mu, \sum)$ can be transformed to $Y = N_p(0, I)$ by an affine transformation

$$Y = A(X - \mu) \tag{10}$$

where A is any $p \times p$ matrix satisfying the condition $A \sum A' = I$. The choice of A is not unique. One choice we made in Section 2 of the paper is $A = \sum^{-1/2}$, a symmetric inverse square root of $\sum$. Other choices can be made such as $A = \Lambda^{-1} P'$ where Λ and P are the matrices appearing in the spectral decomposition of $\sum = P\Lambda^2 P'$. Since μ and $\sum$ are not known, we estimate them by $\bar{X}$ and S from the sample and compute

$$Y_i = \Lambda^{-1} P'(X_i - \bar{X}) \tag{11}$$

where $P\Lambda^2 P'$ is the spectral decomposition of S, using the same notation for sample estimates of P and Λ. (See Srivastava and Hui, 1987). We may then consider the $N = np$ components of the vector

$$Y' = (Y_1', \cdots, Y_n') \tag{12}$$

and apply the UVN test as indicated in Section 2. The test based on the N observations (12) is denoted by T_2 in Tables 1, 2, and 3.

Table 1 gives the p-values for the tests based on T_1 and T_2 on Fisher's data on Iris Setosa with $p = 4$ and $n = 50$. The tests were carried out on the original observations and also after transforming one of the measurements (petal width to logarithms).

	Original Measurements		After Transformation	
	T_1	T_2	T_1	T_2
Skewness $\sqrt{b_1}$	0.0674	0.2471	0.9937	0.8681
Kurtosis (b_2)	0.0311	0.0938	0.1453	0.1829
Combined	0.0184	0.1258	0.3462	0.4062

Table 1: p-values for tests T_1 and T_2 on Iris data ($n = 50, p = 4, N = 200$).

Table 2 gives the p-values for tests based on T_1 and T_2 on the angular data on English and Naqada skulls. Only two angles are considered to test for bivariate normality. Table 3 refers to haematology data considered by Royston (1983) in his discussion on tests of normality.

The test proposed in the paper, especially the one based on T_1 is simple to apply and is likely to be more sensitive to departures from MVN. The test can be included in the software for multivariate methods as an initial test to assess the suitability of MVN as a stochastic model for inferential analysis on observed data.

Studies on power comparisons of T_1 with other tests for MVN and some theoretical investigations on the convergence of the sequence of random variables $Y_i, i = 1, \ldots, n$, defined in (5) will be reported in a future communication.

4 Normal probability plot

The tests described in (7), (8) and (9) in Section 2 are specially designed to test for skewness and kurtosis in a more revealing way than the corresponding tests of Mardia (1980). We recommend their use in a routine way in exploratory multivariate analysis, as in the case of univariate data analysis. Other tests may also be used if they are more appropriate in any given situation to detect nonnormality. In addition, we suggest a normal probability plot of the np observations (see D'Agostino et al, 1990, for details). The graph will be approximately a straight line if the underlying distribution is MVN. Deviations from linearity correspond to various type of nonnormality. Some of these reflect features such as skewness and / or kurtosis. Others reflect features such as the presence of outliers, mixtures in data, and so on. Typical shapes of graphs for different types of departures are given in D'Agostino et al (1990).

Acknowledgments: This research work was supported by the Army Research Office under Grant DAAHO4-98-1-0082.

	English data		Naqada data	
	T_1	T_2	T_1	T_2
Skewness $\sqrt{b_1}$	0.8942	0.8077	0.1429	0.2124
Kurtosis (b_2)	0.4587	0.6751	0.5735	0.7863
Combined	0.7532	0.8892	0.2917	0.4430

Table 2: p-values for tests T_1 and T_2 on skull measurements Aitchinson (1986, p. 385) ($n = 51, p = 2, N = 102$).

Résumé: La littérature statistique présente de nombreuses méthodes pour tester si des données observées proviennent d'une distribution normale multivariée avec un vecteur moyen et une matrice de covariance inconnus. Soit $X_1, \ldots, X_n$ un échantillon iid de taille n d'une distribution normale p-variée. Notons $\bar{X}$ le vecteur moyen et S la matrice de variance-covariance de l'échantillon. La plupart des tests de normalité multivariée sont basés sur le résultat que les $Y_i = S^{-1/2}(X_i - \bar{X}), i = 1, \ldots, n$ sont asymptotiquement iid normal p-variés avec un vecteur moyen nul et une matrice de covariance identité. Les tests développés par Andrews et al., Mardia et d'autres sont des fonctions directes de Y_i. Notons que les np composantes de tous les Y_i mis ensemble peuvent être considérés comme un échantillon asymptotique iid de taille $N = np$ d'une distribution normale univariée avec moyenne 0 et variance 1. Nous testons cette hypothèse à l'aide de quelques tests de normalité univariés basés sur N observations indépendantes. En particulier, nous suggérons les tests univariés basés sur l'asymétrie et l'aplatissement qui sont très sensibles aux déviations par rapport à la normalité, ainsi que des diagrammes de probabilités associés à la loi normale.

Mots clés: Tests de normalité multivariée, diagrammes de probabilités associés à la loi normale.

References

[1] Aitchinson, J. (1986). *The Statistical Analysis of Compositional Data*, Chapman and Hall.

[2] Andrews, D.F., Gnanadesikan, R. and Warner, J.L. (1971). Transformations of multivariate data. *Biometrics* 27, 825–840.

[3] Baringhaus, L. and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika* 35, 339–348.

[4] Cox, D.R. and Small, N.J. (1978). Testing multivariate normality. *Biometrika* 65, 263–272.

	Measurements on variables 1, 2, 5		Measurements on variables 2, 3, 5	
	T_1	T_2	T_1	T_2
Skewness $\sqrt{b_1}$	0.0090	0.0087	0.0404	0.0636
Kurtosis (b_2)	0.0632	0.0267	0.2773	0.3369
Combined	0.0059	0.0027	0.0678	0.1129

Table 3: p-values for tests T_1 and T_2 on haematology data Royston (1983) ($n = 103, p = 3, N = 309$).

[5] D'Agostino, Ralph B., Belanger, A. and D'Agostino, Ralph B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician* 44, 316–321.

[6] Epps, T.W. and Pulley, L.B. (1983). A test for normality based on the empirical characteristic function. *Biometrika* 70, 723–726.

[7] Henze, N. and Wanger, T. (1997). A new approach to the BHEP tests for multivariate normality. *J. Multivariate Analysis* 62, 1–23.

[8] Jobson, J.D. (1991). *Applied Multivariate Data Analysis*, Vol. II, Springer-Verlag, New York.

[9] Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* 36, 149–176.

[10] Puri, M.L. and Rao, C.R. (1975). Augumenting Wilk-Shapiro test for normality. Volume dedicated to A. Linder, 129–139.

[11] Mardia, K.V. (1980). Tests of univariate and multivariate normality. In Handbook of Statistics, Vol. 1, (Ed. P.R. Krishnaiah), 279–320.

[12] Rao, C.R. (1997). *Statistics and Truth: Putting Chance to Work* (Second Edition), World Scientific, Singapore and New Jersey.

[13] Royston, J.P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* 31, 115–124.

[14] Royston, J.P. (1983). Some techniques of assessing multivariate normality based on Shapiro-Wilk W. *Applied Statistics* 32, 121–133.

[15] Shapiro, S.S. and Francia, R.S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association* 67, 215–216.

[16] Shapiro, S.S. and Wilk, S.S. (1968). Approximations to the null distribution for the W statistic. *Technometrics* 10, 861–866.

[17] Srivastava, M.S. and Hui, T.K. (1987). On assessing multivariate normality based on Shapiro-Wilk statistic. *Statistics and Probability Letters* 5, 11–18.

Student (1998), Vol.2, No.4, 325-332

Sur la complémentarité entre analyse factorielle et classification automatique

Fabienne Benninghoff

Centre Interfacultaire de Gérontologie, Université de Genève, Suisse

Résumé: Les procédures d'analyses statistiques multidimensionnelles constituent des stratégies efficaces de dépouillement d'enquêtes. Nous les avons utilisées pour analyser un ensemble de données du microrecensement de l'Office Fédéral de la Statistique portant sur les loisirs des résidents suisses. Notre objectif était de dégager des structures de loisirs à l'aide de l'analyse des correspondances multiples et des profils d'activités de loisirs au moyen de la classification automatique effectuée sur les coordonnées des premiers axes factoriels.

Mots clés: Analyse des correspondances multiples, dépouillement d'enquête, classification automatique.

1 Introduction

Le dépouillement d'enquêtes se fait traditionnellement à l'aide de techniques descriptives relativement simples, à savoir les tris et les tableaux croisés. Cette démarche peut devenir fastidieuse selon le nombre de variables mis en jeu. Or il existe des procédures multidimensionnelles permettant de synthétiser l'information contenue dans de grande base de données (voir par exemple Manly, 1986). Cet article se propose de montrer la richesse et la complémentarité de deux de ces méthodes exploratoires: l'analyse factorielle et la classification automatique (voir Lebart et Salem, 1988).

Pour illustrer notre démarche nous avons utilisé le microrecensement de l'Office Fédéral de la Statistique de 1988 portant sur les loisirs des Suisses. Notre objectif est de comprendre comment se structurent les activités de loisirs, c'est-à-dire de découvrir quelles activités de loisirs sont associées, de dégager des profils de loisirs à partir de ces associations, et de voir

si ces profils sont associés à des groupes d'individus se trouvant dans la même phase de la vie ou partageant la même structure familiale et socio-professionnelle.

2 Données disponibles

La base de données provient du microrecensement suisse "Loisirs et culture" réalisé par l'Office Fédéral de la Statistique en 1988 auprès d'un échantillon de 48 435 résidents suisses âgés de 15 ans et plus. L'objectif de cette enquête était de "saisir les points communs, ainsi que la pluralité de la population suisse dans son comportement culturel en matière de loisirs." (voir OFS, 1992). Grâce au mode d'échantillonnage utilisé, à savoir aléatoire et stratifié, la représentativité de la population suisse est assurée. Dans cet article cependant, seuls les résultats concernant les 13 789 individus âgés de 55 ans et plus sont présentés. Cette sous-population est particulièrement intéressante, car elle permet de capter l'enjeu et l'impact de la retraite sur la structure du temps libre.

La liste des variables utilisées se trouve dans la Table 1. Elle comprend des *activités de loisirs* selon le lieu de réalisation et la fréquence (au moins une fois par jour, par semaine, par mois, par année et jamais), ainsi que des variables liées au *parcours de vie* et au *milieu social.*

3 Techniques d'analyse

L'idée de compléter les opérations de réduction et de visualisation d'information obtenues par l'analyse factorielle, avec des procédures de regroupement réalisées par la classification automatique, est développée dans le logiciel SPAD (Système Portable pour l'Analyse de Données, voir Lebart et al., 1993). Cet enchaînement de méthodes est résumé dans la notion de "thèmascope" et est constitué de six étapes (Lebart, 1989; Morineau, 1993). Les choix que nous avons effectués pour notre analyse sont spécifiés *en italique* pour chaque étape:

1. Choix d'un thème actif: *activités de loisirs.*
2. Analyse factorielle à partir des données actives: *analyse des correspondances multiples.*
3. Positionnement des variables illustratives sur les plans factoriels: *parcours de vie et milieu social.*
4. Classification basée sur les coordonnées factorielles: *algorithme mixte.*
5. Description statistique de chaque classe.
6. Positionnement des centres de classe dans les plans factoriels.

Var. actives	Activités de loisirs	Libellés*
Relations amicales		
à domicile	être avec des amis	**DA.**
hors de chez soi	être avec des amis	**AMI.**
	rencontrer des gens	**REN**
	boire et manger	**REP.**
Divertissements		
à domicile	se distraire (radio, tv, lecture)	**DD.**
	se reposer, se délasser	**DR.**
hors de chez soi	cinéma, discothèque	**CIN.**
	manifestations publiques	**MAN.**
	fêtes	**FET.**
Hobbies		
à domicile	activités manuelles, bricolages	**DB.**
	soins aux plantes, animaux	**DS.**
Activités physiques	activité physique	**SPO.**
	activité en plein air	**NAT.**
Evasion, exploration	voyages, excursions	**VOY.**
Activités culturelles	élargir ses connaissances	**CON.**
	suivre des cours	**COU.**
	activités culturelles	**CUL.**
Spectacles culturels	spectacles culturels classiques	**CLA.**
	événements culturels	**JEU.**
	manifestations culturelles	**POP.**
Engagements	activités religieuses	**REL.**
	activités politiques	**POL.**
	activités associatives	**CLU.**
Var. illustratives	**Modalités**	**Libellés**
Parcours de vie		
Age	groupes d'âge quiquénaux	
Sexe	homme	**H**
	femme	**F**
Structure du ménage	seul	**seul**
	en couple sans enfant	**cple**
	en couple avec enfant	**cp+e**
	seul avec enfant	**sl+e**
	chez enfant	**co-e**
Enfant le plus jeune	groupes d'âge	
Etat matrimonial	célibataire	**celi**
	marié	**mari**
	divorcé	**divo**
	veuf	**veuf**
Activité exercée	actif à temps complet	**tc**
	actif à temps partiel	**tp**
	au foyer	**foy**
	retraité	**rent**
Milieu social		
Niveau de formation	bas [école obligatoire]	**bas**
	moyen [apprentissage]	**moy**
	supérieur	**sup**
Catégorie socio-professionnelle	indépendant	**ind**
	cadre supérieur	**cs**
	cadre moyen supérieur	**cms**
	cadre moyen inférieur	**cmi**
	ouvrier qualifié	**oq**
	ouvrier non-qualifié	**onq**

* Les points laissent place aux *fréquences de l'activité:*

.j = au moins 1x/jour **.s** = 1x/semaine **.m** = 1x/mois
.a = 1x/année **.r** = rarement-jamais **.nr** = non-réponse

Table 1: Liste des variables utilisées.

4 Résultats

L'analyse "thèmascope" réalisée à l'aide du logiciel SPAD est très riche. Nous nous contentons d'en présenter les principaux résultats.

4.1 Analyse des correspondances multiples

Le premier axe factoriel oppose les **répondants** aux **non-répondants** (voir Figure 1). En effet, les individus n'ayant pas répondu aux questions portant sur les loisirs se regroupent à gauche du graphique, ainsi que les individus pratiquant quotidiennement certaines activités telles que les voyages, les fêtes, la religion, les spectacles culturels, les relations amicales, etc. Sur l'autre extrémité du premier axe, se situent les modalités de loisirs pratiqués plutôt rarement (cadran du bas à droite: les activités culturelles, politique et sociales) ou plusieurs fois par an (cadran du haut à droite).

Le deuxième axe montre l'importance de la pratique des loisirs. Les **activités pratiquées rarement voire jamais** - plus particulièrement les relations amicales, les voyages-excursions, les activité en nature - se rassemblent en bas à droite du graphique. Tandis que les **activités de loisirs pratiquées régulièrement** se situent à l'autre extrémité de l'axe. En haut à droite, nous trouvons à nouveau les activités culturelles, politique et sociales mais pratiquées au moins une fois par mois ou par semaine. Remarquons que les modalités notées avec une * ont en réalité des coordonnées plus élevées sur l'axe 2 et ont été ramenées sur le bord du graphe.

Les **variables illustratives** caractérisent davantage les cadrans que les axes (voir Figure 2). Dans le cadran des non-répondants (en bas à gauche), nous trouvons les points moyens des individus les plus âgés (75 ans et plus), vivant seuls, veufs et n'exerçant plus d'activité professionnelle. Dans le cadran opposé des répondants (en haut à droite), sont positionnées les points moyens des individus les plus jeunes (55-59 ans), mariés, vivant en couple avec enfant(s) ainsi que celui des personnes encore insérées à temps complet dans le marché de l'emploi. A ces deux cadrans, qui opposent les grandes phases de la seconde moitié de la vie, est aussi associé le niveau de formation: les générations les plus jeunes présentant un bagage culturel plus élevé.

4.2 Classification basée sur les coordonnées factorielles

A partir des coordonnées sur les 14 premiers axes factoriels, nous avons effectué une classification basée sur l'algorithme mixte proposé dans le logiciel SPAD. Après inspection visuelle du dendrogramme, la partition en 5 classes a été retenue. La Table 2 nous donne une description de ces 5 classes dont les centres ont été positionné sur les plans factoriels (voir Figure 2).

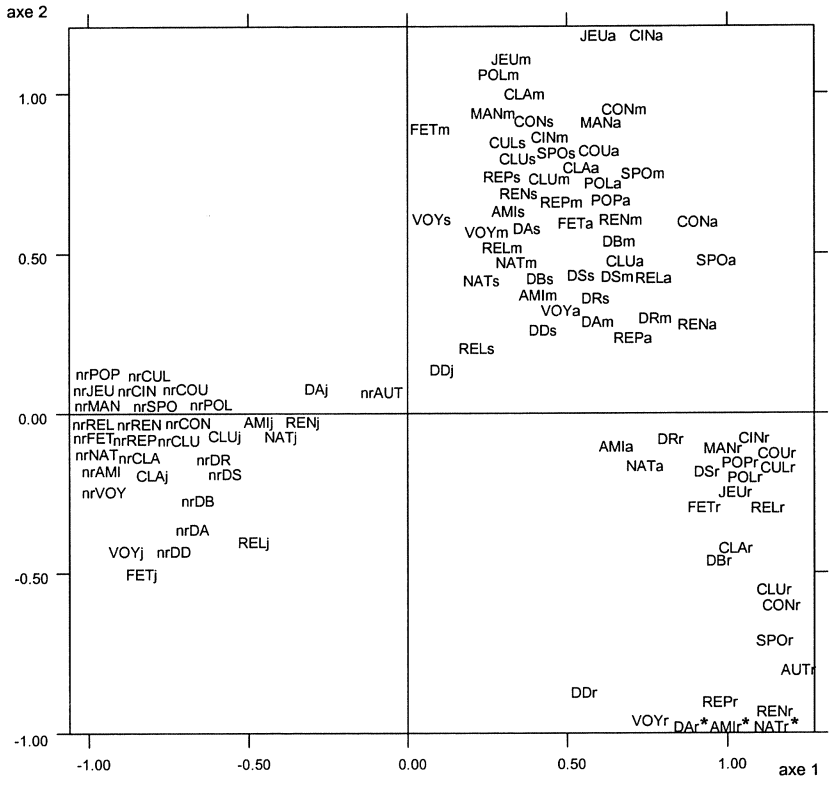


Figure 1: Analyse factorielle de correspondance: variables actives.

La *classe A* se caractérise à la fois par **des activités pratiquées rarement et des activités pratiquées régulièrement**. Cette classe, qui valorise les relations amicales, les activités physiques, la vie associative mais qui s'investit peu dans la vie culturelle, politique et religieuse de ce pays compte proportionnellement plus d'individus dans la partie moyenne et supérieure de l'échelle sociale. Les membres de cette classe sont un peu plus jeunes et davantage insérés dans le marché de l'emploi. Ils sont également plus nombreux à vivre en couple avec ou sans enfant.

La *classe B* présente une structure de loisirs qui se distingue des autres classes par sa relative simplicité: la **rareté des pratiques de loisirs**, particulièrement dans les activités culturelles, sportives, associatives et politiques. Ce profil de loisirs n'est pas associé à un type précis d'individus, bien que nous observons dans cette classe proportionnellement un peu plus de Suisses allemands, de Suisses italiens ainsi que d'individus de bas niveau de formation.

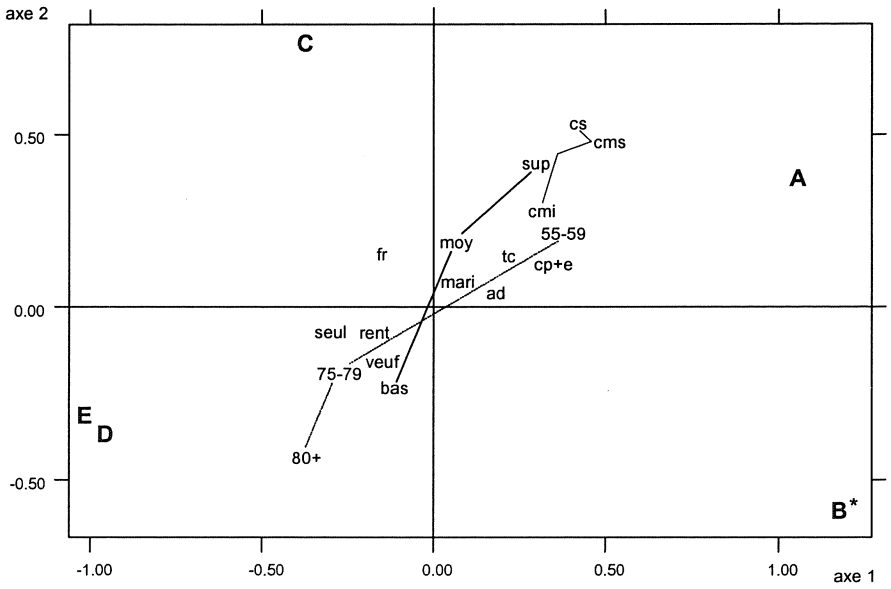


Figure 2: Analyse factorielle de correspondance: variables illustratives.

La *classe C* se caractérise d'une part par des **taux élevés de non-réponses** pour des activités de type culturelles, politiques, etc., ainsi que par des **taux importants de pratique régulière** dans les relations amicales, les activités en nature. Cette classe ne représente pas non plus un groupe spécifique d'individus bien qu'elle contienne légèrement plus d'individus âgés de 65 à 69 ans et de rentiers.

La *classe D* présente un profil de loisirs particulier basé à la fois sur **des activités pratiquées quotidiennement** (relations amicales, voyages-excursions, fêtes, nature, clubs, religion, etc.) et sur des **non-réponses** (activités culturelles, politique, grandes manifestations, sport, etc.) Cette structure de loisirs compte davantage de personnes vivant seules, de retraités, de veufs ainsi que d'individus âgés de 70 ans et plus. Une proportion légèrement supérieure d'individus n'a pas poursuivi de formation au delà de l'école obligatoire.

La *classe E* se distingue de l'ensemble des autres classes uniquement par des **taux de non-réponse très élevés** pour beaucoup d'activités. A cette classe, sont associés les plus âgés. Comme dans la classe précédente, les personnes à la retraite, vivant seules et ayant un faible capital scolaire sont surreprésentés.

Classe %	Variables actives: activités de loisirs	% classe	% global	variables illustratives: caractéristiques	% classe	% global
	- activités rares:			- milieu social:		
	cinéma	85	40	**formation supérieure**	28	18
A	cours	79	37	**formation moyenne**	41	34
	activité culturelle	78	36	- parcours de vie:		
28%	politique, etc.	72	35	**55-59 ans**	35	24
	- activités régulières:			**temps complet**	37	28
	repas 1xmois	33	17	**marié**	77	69
	sport 1xsem.	32	16	**couple avec enfant**	23	17
	rencontre 1xmois, etc.	23	10	présence enfant 19 +	20	15
	- activités rares:			- autres:		
	culture jeune	96	37	suisse allemand	74	68
B	sport	78	21	suisse italien	9	6
	connaissance	79	23	- milieu social:		
15%	cinéma, disco	96	40	formation obligatoire	50	43
	culture classique	87	30	- parcours de vie:		
	clubs	78	24	couple avec enfant	22	17
	activité cult., etc.	91	36	chez enfant	6	3
	- activités manquantes:			- parcours de vie:		
	cinéma, disco	90	55	65-69 ans	23	19
C	cours	90	57	rentier	56	52
	activité cult., etc.	89	55			
23%	- activités régulières:			- milieu social: -		
	ami à l'ext. 1xsem	40	27			
	nature 1xsem., etc.	40	29			
	- activités fréquentes:			- parcours de vie:		
	voyage-excursion	60	5	**seul**	53	27
D	amis à l'extérieur	65	10	**rentier**	72	52
	rencontre	53	8	**veuf**	31	17
6%	fêtes, etc.	31	2	**75-79 ans**	22	11
	-activités manquantes:			70-74 ans	22	14
	cinéma, disco	93	55	80 ans et plus	14	9
	spectacle cult. jeune	90	53	- milieu social:		
	cours, etc.	90	57	formation obligatoire	52	43
	- activités manquantes:			- milieu social:		
	spectacle cult. jeune	99	53	**formation obligatoire**	55	43
E	spect. cult. classique	92	44	- parcours de vie:		
	cinéma, disco.	99	55	**80 ans et plus**	14	9
28%	manifestation	95	49	**rentier**	58	52
	voyage-excursion, etc.	71	27	**seul**	33	27
				75-79 ans	15	11

Table 2: Description des classes obtenues.

5 Conclusion

La stratégie d'analyse exploratoire exposée dans cet article, nous a permis de mieux comprendre les logiques d'organisation des pratiques de loisirs. L'association entre différents types d'activité et leur fréquence de réalisation a permis de mettre en évidence plusieurs profils de loisirs chez les individus âgés de 55 ans et plus. La plus importante césure au niveau des pratiques de loisirs apparaît à cet âge là entre les positions extrêmes du parcours de vie. Au terme de ces analyses exploratoires, tout laisse donc à penser que le parcours de vie est la dimension la plus structurante des pratiques de loisirs, le milieu social étant discriminant essentiellement en matière de non-réponses et d'activités culturelles. Mais, pour montrer le poids réel de chacune de ces variables, il faudrait effectuer d'autres analyses statistiques (régression logistique, modèles log-linéaires, etc.).

Remerciements: Je tiens à remercier le Professeur Lalive d'Epinay et le Professeur Lebart qui m'ont encadré durant ce travail.

Abstract: Multivariate analysis are effective statistical techniques. We have used them for the analysis of data on the spare-time activities of the Swiss residents. Our aim was to exhibit spare-time activities structures and profiles by using correspondence and cluster analysis.

Key words: Data analysis, correspondence analysis, cluster analysis.

Références

[1] Lebart, L. (1989). Stratégies du traitement des données d'enquêtes. Revue Modulad, No 3, Paris.

[2] Lebart, L., Morineau, A., Lambert, T., et Pleuvert, P. (1993). *SPAD, Sytème Portable pour l'Analyse de Données.* Version 2.5, CISIA ed., Paris.

[3] Lebart, L. et Salem, A. (1988). *Analyse de données textuelles.* Dunod, Paris.

[4] Manly, B.F.J. (1986). *Multivariate Statistical Methods: A Primer.* Chapman and Hall.

[5] Morineau, A. (1993). Le "Thèmascope" ou analyse structurelle des données. In *Traitement statistique des enquêtes*, D. Grangé et L. Lebart eds., Dunod, Paris.

[6] OFS (1992). *Loisirs et culture, Microrecensement 1988, Données de base.* Berne.

Student (1998), Vol.2, No.4, 333-336

Sums of Uniforms, Sums of Discrete Uniforms, and the Central Limit Theorem

Peter Thompson

Department of Statistics, North Dakota State University, Fargo, USA

Abstract: The paper demonstrates how exact formulas for probabilities involving sums of independent uniforms and sums of independent discrete uniforms can be used in Central Limit Theorem examples.

Formulas exist for finding exact probabilities concerning sums of independent uniforms and sums of independent discrete uniforms. These can be used by instructors to supplement a treatment of the Central Limit Theorem.

First, we look at sums of independent uniforms. Supose we have $X_1, X_2, \ldots, X_n$, independent uniform (0,1), and $S = \sum_{i=1}^{n} X_i$. Suppose $0 \leq s \leq n$. Then

$$P(S \leq s) = \frac{1}{n!} \sum_{0 \leq k \leq s,\ k \text{ an integer}} \binom{n}{k} (s-k)^n (-1)^k$$

(See Feller 1971, p. 27). This can used to give an "I've seen it with my own eyes!" example of how fast true probabilities involving S approach corresponding approximate normal probabilities.

Example 1 *Even when* $n = 12$, $S-6$ *is very close to a normal (0,1), close enough for computer simulations (Bain and Englehardt, 1991, p. 240). It is easy to demonstrate this to students. For example,*

$$P(S - 6 \leq -0.5) = \frac{1}{12!} \sum_{k=0}^{5} \binom{12}{k} (5.5-k)^{12} (-1)^k = .3106,$$

$$P(normal\ (0,1) \leq -0.5) = .3085,$$

or

$$P(S-6 \leq -1.5) = \frac{1}{12!}\sum_{k=0}^{4}\binom{12}{k}(4.5-k)^{12}(-1)^k = .0.0674,$$

$$P(normal\ (0,1) \leq -1.5) = .0668.$$

Students could be allowed to pick any reasonable number they wanted and a comparison could easily be done in class.

Another application of the formula arises in a seemingly unconnected problem. If $Y_1, Y_2, \ldots, Y_n$ are random variables it is often of interest to look at R=number of times a Y_i exceeds the Y_{i-1} immediately preceding it. If the Y_i are independent from the same continuous distribution, the pdf of R can be shown to be

$$f(r) = \frac{1}{n!}\sum_{k=0}^{r}\binom{n+1}{k}(r+1-k)^n(-1)^k, r = 0,1,\ldots,n-1.$$

The sum in the above expression is called an Eulerian number. See Holte (1997) for a discussion.

In David and Barton (1962) an interesting but involved argument is given to show that R is asymptotically normal. Using our formula, and ideas from Holte (1997), we can take a different approach. Letting R^* denote the greatest integer less than or equal to S, a sum of n uniform (0,1)'s, it is easy to show that

$$P(R^* = r) = \frac{1}{n!}\sum_{k=0}^{r+1}\binom{n}{k}(r+1-k)^n(-1)^k - \frac{1}{n!}\sum_{k=0}^{r}\binom{n}{k}(r-k)^n(-1)^k$$

$$= \frac{1}{n!}\sum_{k=0}^{r}\binom{n+1}{k}(r+1-k)^n(-1)^k.$$

Thus R has the same distribution as R^* and its asymptotic normality follows from the Central Limit Theorem.

Next, we switch to the discrete uniform case. Suppose $X_1, X_2, \ldots, X_n$ are discrete uniform $(1, 2, \ldots, M)$. That is, $P(X_i = k) = 1/M$, $k = 1, 2, \ldots, M$. Let $S = X_1 + X_2 + \ldots, X_n$. Supoose s is an integer, $n \leq s \leq nM$. Then

$$P(S=s) = (\frac{1}{M})^n \sum_{0 \leq k \leq \frac{s-n}{M},\ k \text{ an integer}} \binom{n}{k}\binom{s-1-kM}{n-1}(-1)^k$$

(See Feller, 1968, p. 284).

We can use S in normal approximation problems with continuity corrections. Note that $E(S) = n(M+1)/2$, $Var(S) = n(M^2-1)/12$. In the example we give, we choose n and M so that S has a "convenient" standard deviation, allowing us to hit normal table values exactly and thus avoiding the need for interpolation to demonstrate the accuracy of the normal approximation. Since this is moderately nontrivial, we give a small table of n and M values which would be useful in constructing similar examples.

n	M	SD of S
10	11	10
15	9	10
20	4	5
25	7	10
30	19	30
40	11	20
50	5	10
60	9	20

Example 2 *A 4-sided die (sides 1,2,3,4) is rolled 20 times. Let S=sum of the rolls. Find the normal approximation to $P(S=50), P(S=55), P(S=60)$ and $P(S=65)$ and compare these approximations to the true values.*

Solution:

$$S \approx \text{Normal } (\mu = 50, \sigma^2 = 25)$$

$$\begin{aligned} P(S=50) &= P(49.5 < S < 50.5) \\ &\approx P(-.1 < Z < .1) \\ &= .0796. \end{aligned}$$

Similarly,

$$P(S=55) \approx .0484,$$

$$P(S=60) \approx .0108,$$

$$P(S=65) \approx .0009.$$

The true values are obtained using the formula. We get

$$P(S=50) = \binom{1}{4}^{20} \sum_{k=0}^{7} \binom{20}{k} \binom{49-4k}{19} (-1)^k = .07911.$$

Similarly,

$$P(S = 55) = .04866,$$
$$P(S = 60) = .01096,$$
$$P(S = 65) = .0008064.$$

We give one of final example of a normal approximation. We use the fact that for large n and M odd,

$$P\left(S = \frac{n(M+1)}{2}\right) \approx \frac{1}{\sqrt{2\pi}\sigma}.$$

Example 3 *How many times does a 25-sided die need to be rolled in order to guarantee a less than 1 in a 1000 chance of the sum of the rolls equaling its expected value* $(13n)$*?*

Solution: For large n,

$$P(S = 13n) \approx \frac{1}{\sqrt{2\pi}\sqrt{52n}} \stackrel{\text{set}}{\rightarrow} = .001.$$

We get $n = 3061$. (An exact calculation gives us $P(S = 13(3060)) = 0.00100006$ and $P(S = 13(3061)) = 0.00099990$).

Acknowledgments: The author would like to thank the referee and the editor for the helpful comments they made on an earlier version of this paper.

References

[1] Bain, L. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics.* PWS-Kent, Boston.

[2] David, F. N. and Barton, D. E. (1962). *Combinatorial Chance.* Latimer, Trend and Co., Ltd., Plymouth.

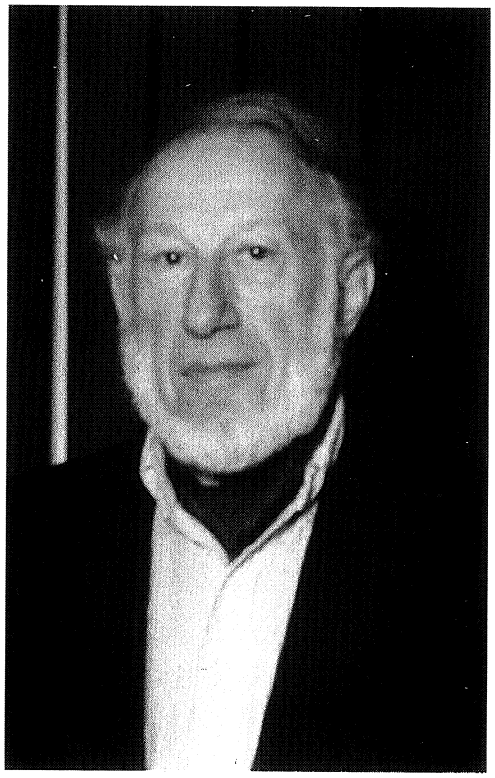
[3] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* Vol. 1, 3^{rd} ed., John Wiley & Sons, New York.

[4] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications.* Vol. 2, 2^{nd} ed., John Wiley & Sons, New York.

[5] Holte, J. (1997). Carries, Combinatories, and an Amazing Matrix. *American Mathematical Monthly*, **107**, 138-149.

Student (1998), Vol.2, No.4, 338-342

Ingram Olkin, Statistical Statesman



Ingram Olkin was born on 23rd July 1924 in Waterbury, Connecticut, USA. He received a Bachelor's degree from the College of the City of New York in 1947, after serving in the army as a meteorologist. He then contined his studies, received a Master's degree in Mathematical Statistics from Columbia University in 1949, and a Ph.D. in Mathematical Statistics from the University of North Carolina in 1951. He taught at Michigan State University and the University of Minnesota before assuming from 1961 his present position at Stanford University as Professor of Statistics and Education.

Dr. Olkin has been an editor of *The Annals of Statistics* and an associate editor of *Psychometrika*, the *Journal of Educational Statistics*, the *Journal of the American Statistical Association*, *Linear Algebra and its Application* and other mathematics and statistical journals. His major interests are in multivariate analysis, inequalities, in the theory and application of meta-analysis and in models in the social, behavioral, and biological sciences. He has co-written a number of books including: *Inequalities Theory of Majorization and its Applications* (1979), *Selecting and Ordering Populations* (1980), *Probability Models and Applications* (1980), and most recently, *Statistical Methods in Meta-Analysis* (1985). He has also served as chairman of the Committee of Applied and Theoretical Statistics, National Research

Council, National Academy of Sciences; chairman of the Special Interest Group in Educational Statistics; President of the Institute of Mathematical Statistics, and Chair, Committee of Presidents of Statistical Societies. He is currently a member of the Technical Advisory Committee for the National Assessment of Educational Progress and a member of the Board of Trustees, National Institute of Statistical Sciences. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics. He is a recipient of the Wilks Medal, and was awarded an honorary Doctor of Science degree by de Montfort University.

In "A Conversation with Ingram Olkin" (published in Gleser, L.J. et al. Eds. (1989), *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin,* Springer, pp. 7-33), Ingram Olkin answers to the question "What is your assessment of the current state of the health of the field of statistics, and where do you see the field heading?" as follows:

"I am a bit worried about statistics as a field. As you know, I come from the mathematical community and I've always liked the mathematics of statistics. But I think that the connection with applications is an essential ingredient at this time. I say that because applications are crying out for statistical help. We currently produce approximately 300 Ph.D.'s in statistics and probability per year. This is a small number considering the number of fields of application that need statisticians. Fields such as geostatistics, psychometrics, education, social science statistics, newer fields such as chemometrics and legal statistics generate a tremendous need that we are not fulfilling. Inevitably this will mean that others will fulfill those needs. If that happens across fields of application, we will be left primarily with the mathematical part of statistics, and the applied parts will be carried out by others not well-versed in statistics. Indeed, I think that a large amount of statistics is now being carried out by non-statisticians who learn their statistics from computer packages and from short courses. So I worry about this separation between theory and practice and the fact that we are not producing the number of doctorates to fulfill needs in all of these other areas".

He then adds: "There is still an excitement in the field, but my impression is that, except for a few places, the growth in statistics departments has reached a plateau. I believe that this is true because we do not have a natural mechanism for statistics departments to create strong links to other departments of the academic community. Academic institutions have not been designed for cross-disciplinary research, and indeed may actually be antagonistic to cross-disciplinary research".

Memories from the International Statistical Institute Conference of 1961 in Paris.

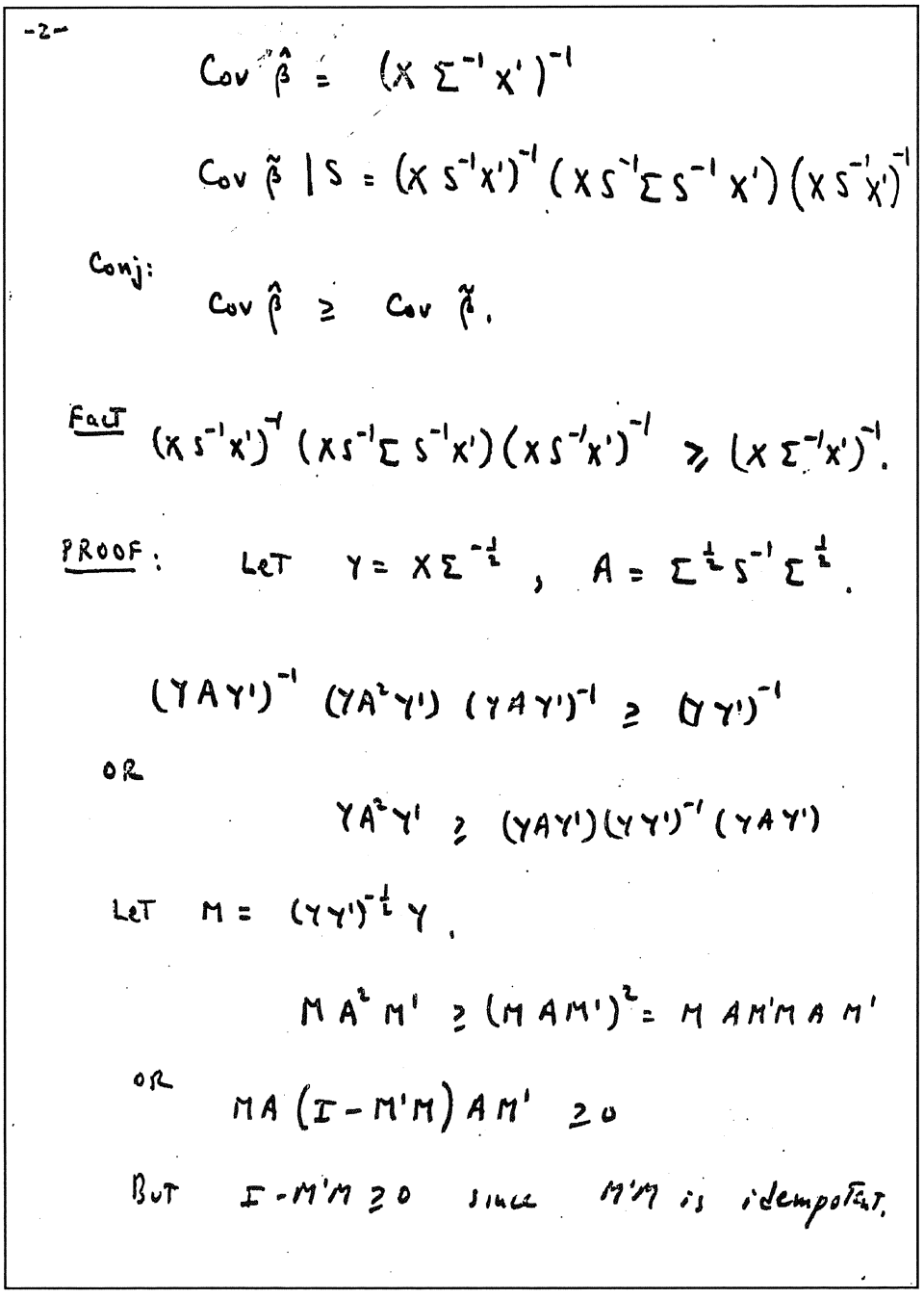


From left to right, Sam Greenhouse, R.A. Fisher, an unknown participant, Carol Parzen, Ingram Olkin and Manny Parzen.

Memories from the International Statistical Institute Conference of 1961 in Paris.



From bottom to top are on the left side of the table, an unknown participant, Elizabeth Scott, Jerzey Neyman, Anne Durbin, Jim Durbin, Miriam Chernoff and Herman Chernoff. From bottom to top are on the right side of the table, Jack Youden, Ingram Olkin, Dorothy Gilford, Manny Parzen, Carol Parzen, Ellen Chernoff and Judy Chernoff.



Olkin's handwriting showing that one estimator is preferable to another by showing that the covariance matrices are ordered (from class notes about 1968).

Student (1998), Vol.2, No.4, 343-348

Why is Matrix Analysis Part of the Statistics Curriculum

Ingram Olkin

Stanford University, USA

Abstract: We provide a discussion of several important areas of intersection between matrix theory and statistics.

Key words: Matrix factorizations, extremal problems, multivariate distributions, inequalities.

1 Introduction

The Masters program at the Université de Neuchâtel lists only one mathematics course "Matrix Theory and Inequalities". One might argue for the inclusion of other mathematics courses, but it is hard to argue that this topic does not play a central role in many statistical contexts. Matrix analysis is part of a warehouse of handy mathematical tools.

In this note I review some areas in which matrix theory (including inequalities) have natural statistical origins:

1. Matrix factorizations,
2. Extremal problems,
3. Multivariate integrals,
4. Inequalities,
5. Independence properties.

Each of these topics is large, and we provide only a few examples to illustrate the intersection of statistics and matrix analysis.

2 Matrix Factorizations

This is a topic that is currently in vogue, primarily because factorizations

often help to reduce a problem to a simpler form. Only a few factorizations suffice for most statistical purposes. In the following all matrices are real; there are counterpart factorizations for complex matrices.

Theorem 1 *Every $p \times p$ symmetric matrix S can be factored as*

$$S = GDG',$$

where G is orthogonal, $D = diag(\lambda_1, ..., \lambda_p)$, and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ are the ordered eigenvalues of S.

Theorem 2 *Every $p \times p$ positive semi-definite matrix S can be factored as*

$$S = LL',$$

where $L = (l_{ij})$ is lower triangular with $l_{ii} \geq 0$.

The matrix L is sometimes called the triangular square root of S, in contrast to the symmetric positive semi-definite square root V in $S = V^2$. In Theorem 1 if $S = GDG'$, then $V = GD^{1/2}G'$.

Theorem 3 *Every $p \times p$ matrix A can be factored as*

$$A = GDH,$$

where G and H are orthogonal, $D = diag(\alpha_1, ..., \alpha_p)$, and $\alpha_1 \geq ... \geq \alpha_p \geq 0$ are the ordered singular values of A, that is

$$\alpha_i = \lambda_i^{1/2}(AA'), \quad i = 1, ..., p.$$

Other factorizations arise from numerical analytic problems. For example, every $p \times p$ matrix $A = QR$, where Q is lower triangular with $q_{ii} = 1$, and R is upper triangular.

3 Extremal Problems

The method of maximum likelihood is a basic concept in the estimation of the parameters of a distribution. In univariate analysis the maximizations can often be carried out with standard calculus methods. However, in multivariate analysis we almost always are confronted with a matrix problem, as in the following:

$$\max_{\Omega} (\det \Sigma)^{-n/2} \exp\left[-\frac{1}{2} tr\left(\Sigma^{-1} S\right)\right], \tag{1}$$

where $\Omega = \{\Sigma : \Sigma \text{ is positive definite}\}$ and S is positive definite.

To simplify this problem note that $tr\left(\Sigma^{-1}S\right) = tr\left(V\Sigma^{-1}V\right)$, where $V = S^{1/2}$. With $\Psi = V\Sigma^{-1}V$ (1) becomes (except for a constant)

$$\max_{\Omega_1} \left(\det \Psi\right)^{-n/2} \exp\left[-\frac{1}{2} tr\ \Psi\right], \tag{2}$$

where now $\Omega_1 = \{\Psi : \Psi \text{ is positive definite}\}$.

Using the factorization of Theorem 1: $\Psi = GDG'$, where G is orthogonal and $D = diag\left(\lambda_1, ..., \lambda_p\right)$ and the λ_i are the eigenvalues of Ψ we obtain from (2):

$$\max_{\lambda_i > 0} \left(\prod_1^p \lambda_i\right)^{-n/2} \exp\left[-\frac{1}{2}\sum_1^p \lambda_i\right] = \prod_1^p \max_{\lambda_i > 0} \left(\lambda_i^{-n/2} e^{-\frac{1}{2}\lambda_i}\right).$$

We have now reduced the multivariate problem to a product of univariate problems, for which the solution is $\widehat{\lambda}_i = n$. Consequently, the maximum likelihood estimator $\widehat{\Psi}$ of Ψ is

$$\widehat{\Psi} = G\left(nI\right)G' = nI,$$

from which

$$\widehat{\Sigma} = V\widehat{\Psi}^{-1}V = V^2/n = S/n.$$

This result is the multivariate version of the maximum likelihood estimator of the population variance.

4 Multivariate integrals

The normal distribution and the chi-square distribution are central distributions in univariate analysis. The multivariate counterparts are the multivariate normal distribution and the Wishart distribution. The latter generates an integral that we need to evaluate:

$$\int_\Omega \left(\det S\right)^{(n-p-1)/2} \exp\left[-\frac{1}{2} tr\ S\right] \prod_{i<j} ds_{ij}, \tag{3}$$

where $\Omega = \{S : S \text{ is positive definite}\}$.

To simplify this problem we use the factorization of Theorem 2: $S = LL'$. Then

$$\det S = \prod l_{ii}^2, \quad tr\ S = \sum_{i \le j} l_{ij}^2.$$

Consequently, the integral (3) becomes

$$\int_\omega \prod l_{ii}^{n-p-1} \exp\left[-\frac{1}{2}\sum_{i \le j} l_{ij}^2\right] J \prod_{i \le j} dl_{ij}, \tag{4}$$

where $\omega = \{l_{ij} : l_{ii} > 0, -\infty < l_{ij}\,(i \neq j) < \infty\}$, and J is the Jacobian of the transformation involved in Theorem 2. We do not provide the details of the computation of

$$J = 2^p \prod_1^p l_{ii}^{p-i+1},$$

but urge the reader to verify this result for $p = 3$ or 4. Using this value of J the integral (4) reduces to a product of integrals

$$2^p \prod_1^p \left\{ \int_{l_{ii}>0} l_{ii}^{n-i} e^{-\frac{1}{2}l_{ii}^2} dl_{ii} \right\} \left(\int_{-\infty<l_{ij}<\infty} e^{-\frac{1}{2}l_{ij}^2} dl_{ij} \right)^{p(p-1)/2}, \qquad (5)$$

where now each integral is a univariate integral. As

$$\int_0^\infty x^{n-i} e^{-\frac{1}{2}x^2} dx = 2^{\frac{n-i+1}{2}} \Gamma\left(\frac{n-i+1}{2}\right)$$

and

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi},$$

the integral (5) is equal to

$$2^{pn/2} \pi^{p(p-1)/4} \prod_1^p \Gamma\left(\frac{n-i+1}{2}\right).$$

5 Inequalities

The Cauchy-Schwarz inequality is a cornerstone inequality that arises under a variety of guises. For vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ the C-S inequality is

$$(xy')^2 \leq (xx')\,(yy'). \qquad (6)$$

If we let $x = uA^{1/2}$ and $y = vA^{-1/2}$, where A is positive definite, then (6) becomes

$$(uv')^2 \leq (uAu')\left(vA^{-1}v'\right).$$

Throughout Inequality (6) has an integral representation

$$\left(\int f(x)\, g(y)\, dx\, dy\right)^2 \leq \int f^2(x)\, dx \int g^2(y)\, dy \qquad (7)$$

from which we obtain that the squared correlation $\rho^2(x, y)$ between the random variables x and y is less than or equal to 1.

A more general inequality than (7) is the Hölder inequality

$$\int \prod_1^n f_i^{q_i}(x)\,dx \leq \prod_1^n \left[\int f_i(x)\,dx\right]^{q_i}, \quad q_i \geq 0, \quad \sum q_i = 1. \tag{8}$$

In (7) and (8) we assume finite integrability.

A discussion of inequalities would be incomplete without a mention of the arithmetic mean-geometric mean inequality

$$\sum_1^n x_i g_i \geq \prod_1^n x_i^{g_i}, \quad x_i \geq 0, \quad g_i \geq 0, \quad \sum g_i = 1.$$

Even when the x_i may be negative, and $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, we have that

$$x_{(1)} \leq \sum_1^n x_i g_i \leq x_{(n)}. \tag{9}$$

In another context, if x is a vector and S is a positive definite matrix, then

$$\lambda_1 \leq \frac{xSx'}{xx'} \leq \lambda_n, \tag{10}$$

where $\lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues of S. Inequality (10) can be obtained by using the factorization of Theorem 1.

The theory of least squares provides a variety of opportunities for the development of inequalities. Given the equation

$$xA = b,$$

if A is invertible, then the solution is $x = bA^{-1}$. However, if x is a $1 \times p$ vector and A is a $p \times n$ matrix, then A does not have an inverse. This leads to the development of a theory of generalized inverses, and permits us to find the vector x_0 for which $x_0 A$ is "closest" to b. That is, the Euclidean distance $\|x_0 A - b\|$ is minimized.

6 Independence properties

If $x_1, \ldots, x_n$ are independent random variables, with a common standard normal distribution, we ask when are two linear forms $L_1 = xa'$ and $L_2 = xb'$ independent. The answer is that a and b must be orthogonal, that is $ab' = 0$.

This suggests that we ask when is a linear form $L = xa'$ independent of a quadratic form $Q = xAx'$, where A is symmetric. Now the answer is that $aA = 0$.

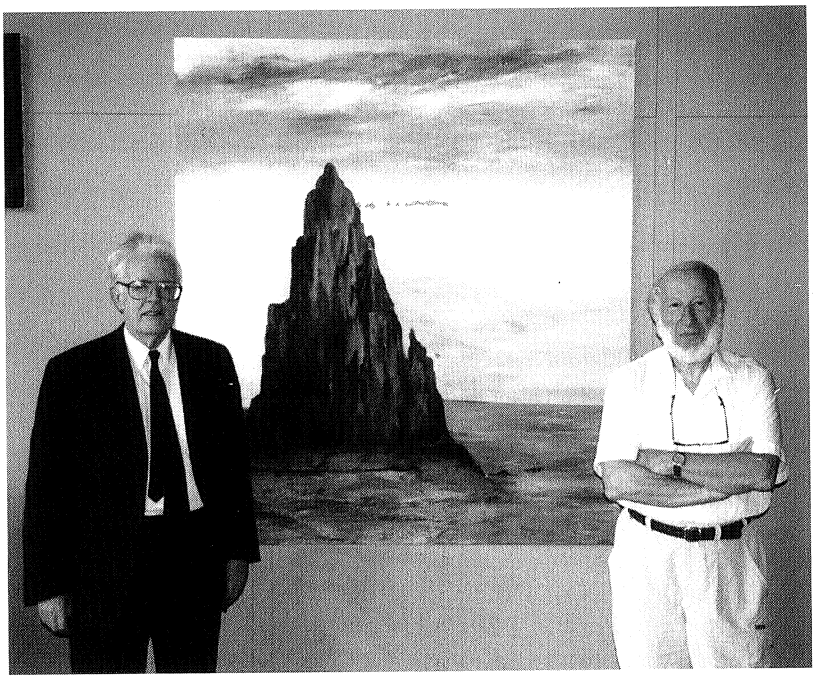
To move to the next step, when are two quadratic forms $Q_1 = xAx'$ and $Q_2 = xBx'$, where A and B are symmetric, independent. The natural conjecture is that $AB = BA = 0$. Indeed, this is the correct answer. But in order to prove this we have an intermediate step:

$$\det(I - \alpha A)\det(I - \beta B) = \det(I - \alpha A - \beta B) \tag{11}$$

must holds for all α and β. The proof that (11) implies $AB = BA = 0$ is not simple. A good exercise is to show this result for the special case $n = 2$.

7 Summary

The move from univariate statistical inference to multivariate statistical inference means that matrices will replace scalars. This in turn generates a wide class of problems. In this short summary we attempt to provide only a few prototype problems that arise in a statistical context. Matrix analysis is a useful tool that permits insights into the geometry of some of the models. It is difficult to imagine a curriculum in which a knowledge of linear algebra and matrix analysis is omitted.



Ingram Olkin with Sir David Cox and a mountain of yet unsolved problems.

Student (1998), Vol.2. No.4, 351-379

Frequency Distribution of Illustrated Scenes in Persian Manuscripts

Farhad Mehran

Statistics Group, University of Neuchâtel, Switzerland

Abstract: Throughout the centuries, the repertoire of Persian miniature painting has almost entirely consisted of illustrations of scenes of great literary texts, the most celebrated of which is the national epic, the Shahnameh. Other vastly illustrated texts are the Khamseh Nizami, the Kalilah wa Dimnah, and the Haft awrang Jami. In each of these works, certain scenes have been illustrated more often than others, and this variation in the recurrence of illustrated scenes could be exploited with advantage, for example, for dating illustrated manuscripts, establishing their genealogy, or reconstructing the cycle of illustrations in dismantled manuscripts. In this paper, we define the concept of a random illustrated manuscript, and derive the probability distribution of scene frequencies in a large sample of manuscripts. We show that a particular distribution, coined the Diffused Waring distribution with probability function $F(x) = x/(a+x)$, closely fits the observed scene frequencies of the Shahnameh and the Haft awrang, and to a lesser degree, the Khamseh and the Kalileh wa Dimnah.

Key words: Random illustrated manuscripts, scene frequencies, Generalized Waring distribution, Diffused Waring distribution.

1 Scene frequencies

1.1 Illustrated Shahnameh scenes

The Shahnameh is an immensely popular poem of about 60'000 couplets composed at the end of the tenth century by the great Iranian poet Firdowsi. Drawing from old oral traditions and ancient Zoroastrian texts, the Shahnameh begins with the creation of the world and narrates the story of

fifty kings in three distinct parts: mythic, epic and historic[1].

The scenes of the Shahnameh have been repeatedly illustrated, on almost every conceivable medium and in virtually all periods. Shahnameh scenes have been depicted on carpets, textiles, tiles, potteries, mina plates, metal vessels, ivory objects, lacquerworks, mural sculptures, oil canvasses and above all, on book leaves for manuscript illustration, album composition and margin decoration, or for single-sheet paintings or drawings. Manuscript illustrations of the Shahnameh exist from the fourteenth century onwards, in most styles of Persian painting (Robinson, 1982), as well as in foreign styles such as Ottoman, Mamluk, Kashmiri, Mughal and pre-Mughal Jaina. There are also nineteenth century illustrated manuscripts in lithographic and printed editions made in Bombay and Tehran (Safynedjad, 1995).

1.2 The Michigan database

The most comprehensive list of Shahnameh manuscript illustrations known to the author has been assembled by Jill Norgren and Edward Davis, at the University of Michigan, under the guidance of Oleg Grabar, Norgren and Davis (1969). Apart from frontispieces, some finispieces and certain unrelated illustrations, the list contains 4122 references to Shahnameh illustrations in more than 100 manuscripts belonging to fifty-five libraries, museums, and private collections throughout the world. Most of these are from Topkapi Saray Museum in Istanbul (967 illustrations), The British Museum in London (534), Bibilothèque Nationale in Paris (281) and The Metropolitan Museum of Art in New York (280). More than half of the illustrations (2512) are from dated manuscripts ranging from 1330 to 1649 with significant concentration in manuscripts of the fifteen and sixteen centuries. About half of the remaining undated illustrations are attributed to specific periods with a ten-year or less margin of uncertainty. Overall there are proportionally more 16th century illustrations than 15th, more 17th century than 14th, and virtually no illustrations of the 18th century or later.

The 4122 illustrations are listed according to the attributed title of the Shahnameh scene they represent, and the list distinguishes 645 such scenes. The complete data are given in Table 8 at the end of this paper. The information contained in the Michigan list has been transferred by the author of the present paper to a diskette as a Paradox table (Paradox for

[1]The most recent and authoritative version of the Shahnameh is the edition by Khaleghi-Motlagh (first volume, 1995). Warner and Warner (1925) is an old English translation and Levy (1987) a more recent prose translation.

Windows, Version 4.5, 1993) and is available upon request. The content has been slightly edited for consistency and obvious misprints, but its accuracy has not been verified against other sources[2].

A summary of the database is presented below as annex of this paper. It gives the titles of the 645 Shahnameh scenes of the Michigan compilation identified by a three-digit serial number, together with the frequency of its illustrations. The list is presented in the order of the unfolding of the Shahnameh stories. Thus, the first Shahnameh illustrated scene is the enthronement of Kayumars (M001) and this scene has been illustrated 49 times according to the Michigan list. The last of illustrated scenes of the Shahnameh (M645) is the killing of Mahuy by Bizhan for avenging the assassination of Yazdgered, the last Sasanid king of Iran. This scene has been illustrated 7 times.

1.3 Incidence matrix

The information on the occurrence of Shahnameh scenes can be conveniently organized into an incidence matrix $\boldsymbol{X}$, where each row, i, of the matrix represents an illustrated Shahnameh manuscript and each column, j, a particular illustrated scene. The value of each cell of the matrix, $\boldsymbol{X}_{ij}$, indicates the presence or absence of the scene in the manuscript. Thus, $\boldsymbol{X}_{ij} = 1$ if scene j is illustrated in manuscript i, and $\boldsymbol{X}_{ij} = 0$ if it is not.

The definition of the incidence matrix assumes that the resolution of the scene descriptions is such that an illustrated manuscript cannot contain more than one illustration of the same scene. A Shahnameh manuscript is defined to be any manuscript which contains illustrations of Shahnameh scenes, including incomplete and truncated manuscripts, and albums (Muraqqa) or anthologies. Single-sheet Shahnameh illustrations as well as detached Shahnameh folios of unknown manuscripts are also assimilated as Shahnameh manuscripts.

It is convenient to assign the columns of $\boldsymbol{X}$ in the order of the unfolding of the story, and the rows in the chronological order of the manuscripts. Thus each row of $\boldsymbol{X}$ represents an illustrative cycle of the corresponding manuscript, and each column the pattern of occurrence of the scene through time. The representation of the illustrative cycles of manuscripts in terms of 1 and 0 has a number of advantages. It is compact, simplifies computer processing, has visual appeal and helps detection of patterns and

[2]No attempt has been made to correct the list with respect to its content and coverage. A process has been started by the author of this paper to use the available lists to supplement the Michigan database, and it is estimated that when completed the extanded database will contain references to some 12'000 or more Shahnameh illustrations.

comparison of cycles from different manuscripts. Another advantage is the possibility of positioning special markers (•) in between known illustrations for potential illustrated scenes on lost folios.

The incidence matrix corresponding to the Shahnameh manuscripts of the Michigan database has $m = 111$ rows (manuscripts), $M = 645$ columns (distinct scenes), and $N = 4122$ non-zero entries (illustrations). An extract for the mythic scenes of the Shahnameh in the fourteenth-century manuscripts listed in the Michigan database is reproduced in Table 1.

M001	00101000000000100000000100000000010000
M002	00000000000000000000000000000000000000
M003	10000000000000000000000000001001001000
M004	00000000000000000000000000000000000000
M005	00000000000000000000000000000000000000
M006	00000000000000000000000000001000000000
M007	00000000000010010000110000001001110000
M008	00000000000000000000100000000010000000
M009	00000000000000000000000000000000000000
M010	00000000000000000000000010000000000000
M011	00000000000000000000000000000000010000

Table 1: Extract of the incidence matrix corresponding to mythic scenes in 14th century illustrated Shahnameh manuscripts.

1.4 Range of scene frequencies

The most frequently illustrated scene of the Shahnameh is Rostam's seventh feat when he kills the White Div (S108). This scene has been illustrated 74 times according to the Michigan list. This scene has been among the most commonly illustrated Shahnameh scenes in each of the four centuries represented in the Michigan database. Overall one can expect to find a painting of this scene in about two-third of the illustrated Shahnameh manuscripts.

The popularity of the scene is no doubt related to the episode it depicts. It shows the Iranian great hero Rostam slaying the White Div of Mazandaran, with its feet cut-off while the local ruler and unwilling guide Owlad, tied to a tree, and Rostam's horse, Rakhsh, are both watching. Rostam had to perform seven feats, the last of which was to kill the White Div, who lived in a dark cavern, in order to obtain its blood to heal the sight and rescue the Iranian king, Kay Ka'us, who foolishly had attacked Mazandaran where he was blinded and taken prisoner, with his troops by the forest demons. A magnificent early sixteenth-century depiction of this scene from

the reknown Shah Tahmasb Shahnameh (1524) is shown in Figure 1.



Figure 1: Painting from the reknown Shah Tahmasb Shahnameh (1524).

With one exception, the next five most illustrated scenes of the Shahnameh in order of abundance (the number in parentheses refer to the frequency of the illustrations in the Michigan database) are also scenes in which the national hero, Rostam, appears:

S149 The last battle between Rostam and Sohrab (67)
S157 Fire-ordeal of Siyavosh (67)
S400 Rostam shoots a double-pointed arrow in the eyes of Esfandiyar (66)
S403 Rostam slays Shagad then dies (64)
S253 Rostam pulls the Khaqan of Chin from the elephant by lasso (59)

On the other extreme, there are Shahnameh scenes which are illustrated very rarely. For example, scene S151 in which Sohrab's body is carried to Zabolestan after the dramatic battle between the father Rostam and the son Sohrab is illustrated only once, in the Metropolitan Museum of Art miniature dated 1605-08. The rare scenes have special significance. They provide helpful indications for identifying and dating illustrated manuscripts. The Michigan database contains 204 unique scenes, 117 scenes that have been illustrated exactly twice, and so on. The complete range of scene frequencies is given in Table 2.

Table 2 shows that the distribution of the frequency of Shahnameh scenes is highly skewed: there are many scenes which are illustrated rarely, and a few which are illustrated frequently. In the remaining of the paper, an attempt is made to derive a theoretical function which fits closely the Shahnameh frequencies in Table 2, as well as similar scene frequencies observed for other Persian literary texts.

x	m_x	x	m_x	x	m_x	x	m_x	x	m_x
1	204	16	3	31	3	46	-	61	-
2	117	17	4	32	3	47	-	62	-
3	57	18	5	33	1	48	1	63	-
4	45	19	5	34	-	49	1	64	1
5	29	20	5	35	-	50	-	65	-
6	35	21	1	36	1	51	-	66	-
7	20	22	-	37	-	52	-	67	2
8	13	23	3	38	1	53	-	68	-
9	14	24	5	39	-	54	1	69	-
10	7	25	3	40	-	55	-	70	-
11	11	26	1	41	-	56	-	71	-
12	8	27	3	42	2	57	1	72	-
13	7	28	-	43	-	58	2	73	-
14	8	29	-	44	1	59	1	74	1
15	5	30	2	45	1	60	-	75	-

Table 2: Range of frequencies of illustrated Shahnameh scenes (m_x = number of Shahnameh scenes illustrated exactly x times.

1.5 Average number of illustrations per scene

In Table 2, the number of scenes appearing exactly x times in the database

is denoted by m_x, $x = 1, 2, ..., \max$, where max represents the highest scene frequency. We note

$$M = \sum_{x=1}^{\max} m_x$$

the total number of distinct scenes and

$$N = \sum_{x=1}^{\max} x m_x$$

the total number of illustrations in the database. It can be shown that M increases with N, but at a decreasing rate. While, the theoretical maximum number of illustrated Shahnameh scenes is unknown, it should not conceivably surpass the number of lines in Ferdowsi's poem, approximately, 60,000. In theory, therefore, as the number of Shahnameh illustrations increases indefinitely, the number of distinct scenes should tend to a finite asymptotic value.

The average number of illustrations per scene is the mean of the distribution of the scene frequencies,

$$\mu = \frac{\sum x m_x}{\sum m_x} = \frac{N}{M}.$$

It also represents the average waiting-time for the appearance of a new scene. For the Shahnameh database, it is

$$\mu = \frac{4122}{645} = 6.4$$

i.e., in about every 6 illustrations, one should expect one new Shahnameh scene. The waiting-time is low at the early stage of manuscript illustration, and high at the later stages. Thus, in the fourteenth century, the waiting-time for the appearance of a new Shahnameh scene was about, 1.6; it increased to 7.2 in the fifteenth century, to 12.7 in the sixteenth century and to 14.7 in the seventeenth century. The limiting value is infinite,

$$\lim_{N \to \infty} \mu = \infty.$$

2 Random illustrated manuscripts

2.1 The concept

A random illustrated manuscript, or random manuscript for short, is a sequence of scenes randomly selected for a complete manuscript (Mehran,

1996, 1997, 1998). It can be regarded as a random sample of scenes drawn according to probabilities from a comprehensive list of illustrated scenes of the Shahnameh. It is assumed that the drawings are carried out one at a time, independently from one another, and without replacement. In practice, each scene probability may be determined in terms of the target size of the manuscript (the number of Shahnameh illustrations intended for the manuscript) and the scene frequency in the Shahnameh database, which may vary from one scene to another.

That the sample should be drawn from a "comprehensive list" is to ensure that each possible illustrated Shahnameh scene has a chance to be selected in the random manuscript. The requirement that the sample be drawn "without replacement" is to avoid generating manuscripts with more than one illustration of the same scene. The independence assumption is for the sake of simplicity, but can be relaxed if necessary. The specification that the scene probabilities be dependent on the scene frequencies is to ensure that the random manuscripts are on the whole representative of the range of frequencies of the illustrated scenes in the database. Finally, the idea of specifying a "target" sample size as opposed to a fixed sample size is to conform with the presumed practice that the number of illustrations of manuscripts were not generally set in advance, but decided upon in the course of the preparation of the manuscript and according to the resources available. There is, however, no difficulty in formulating random manuscripts with fixed sample sizes if this is required.

A random manuscript is specified by a vector $\boldsymbol{X}=(X_1, X_2, ..., X_M)$, where each component X_j is a binary random variable taking the values 1 or 0, depending on whether or not the scene j is illustrated in the manuscript: $X_j = 1$ if scene j is in the illustrated manuscript, and $X_j = 0$ if not.

2.2 Scene probabilities

The probability that scene j is illustrated in a random manuscript is given by,

$$\pi_j = \Pr(X_j = 1), \qquad j = 1, \ldots, M, \tag{1}$$

where M is the number of distinct scenes in the universe of illustrated Shahnameh scenes. The probability that scene j is not illustrated is

$$\Pr(X_j = 0) = 1 - \pi_j.$$

The probability of joint occurrence of scenes j and k in the random manuscript is given by

$$\Pr(X_j, X_k = 1) = \pi_{jk}. \tag{2}$$

The assumption of independence of the scenes implies that when j and k are different, $\pi_{jk} = \pi_j \pi_k$, and when $j = k$, $\pi_{jk} = \pi_j = \pi_k$.

The independence assumption may not always hold in practice, groups of Shahnameh scenes such as the battle of the twelve Rokhs or the Haft Khan-e Rostam tend to appear together in certain illustrated manuscripts. Also, there may be Markov dependence among certain successive scenes. For example, the presence of the illustrated scene S031 "Tur and salm slay Iraj" increases the likelihood of the occurrence of the subsequent illustrated scene S032 "Iraj's coffin brought to Feridun". Conversely, if the scene S315 "Key Khosrow battles Shideh" is illustrated in a manuscript, it is unlikely that the manuscript also includes an illustration of the subsequent scene S316 "Key Khosrow throws Shideh on the ground". Departures from independence tend to increase the occurrence of clusters of illustrated scenes in the manuscript. Thus, the appropriate formulation of random manuscripts in such circumstances would be a cluster sample design.

2.3 Size of manuscripts

The size of an illustrated manuscript is the number of illustrations in the manuscript. In a random manuscript, it corresponds to the sample size of illustrated scenes selected for the manuscript. The size of a manuscript is a random variable given by $n = X_1 + X_2 + ... + X_M$, with expectation, $E(n) = \pi_1 + \pi_2 + \ldots + \pi_M$, and variance, $Var(n) = \sum_{jk}(\pi_{jk} - \pi_j \pi_k)$.

The sizes of manuscripts of a given corpus are obtained as row totals of the incidence matrix. The median size of manuscripts in the Michigan database is 30, and the average size 36. The standard deviation of the distribution is 25. The sizes of Shahnameh manuscripts in the Michigan database follow closely a gamma distribution. The parameters estimated by the method of moments, are $a = 1.689907$, for the shape parameter, and $b = 19.38375$, for the scale parameter. A shift parameter $\gamma = 3$ is also added to account for the fact that the data cover only manuscripts with sizes of 3 and above. Table 3 compares the fitted distribution with the observed values. The goodness-of-fit test gives a chi-square value of 14.8, which is not significant at the 5% level for 11 degrees of freedom.

2.4 Scene popularity

The probability of occurrence of a particular scene in an illustrated manuscript is assumed to depend, in its simplest form, on the size of the manuscript and on the popularity of the scene. The probability of occurrence of a scene in a manuscript of size n may, thus, be expressed by

$$\pi = \pi(\theta, n) \tag{3}$$

	Number of manuscripts		
Size of manuscript	Michigan	Gamma	Difference
0-9	13	10.4	2.6
10-19	23	23.9	-0.9
20-29	19	22.1	-3.1
30-39	21	17.1	3.9
40-49	8	12.4	-4.4
50-59	11	8.6	2.4
60-69	4	5.8	-1.8
70-79	2	3.8	-1.8
80-89	4	2.5	1.5
90-99	5	1.6	3.4
100+	1	2.8	-1.8
Total	111	111	$\chi^2 = 14.8$

Table 3: Size of Shahnameh manuscripts: Michigan database vs Gamma distribution.

where θ is an index of the popularity of the scene measured, say, in a scale of 0 to ∞. Everything else equal, the higher the popularity of a scene, the higher the probability of its occurrence in an illustrated. Thus, the first derivative of π with respect to θ is positive,

$$\frac{\partial \pi_\theta}{\partial \theta} > 0.$$

Similarly, the larger the size of the manuscript, the higher the probability of occurrence, irrespective of the popularity of the scene, i.e., the first derivative of π with respect to n, assumed to be continuous, is also positive,

$$\frac{\partial \pi_n}{\partial n} > 0.$$

The following boundary conditions may also be imposed. The probability of the occurence of any scene in a manuscript of infinite size is 1, so is for a scene of infinite popularity, in a manuscript of any size. Thus, $\pi(\theta, \infty) = 1$ and $\pi(\infty, n) = 1$. At the other extreme, one may assume that a scene with zero popularity has zero probability of occurrence in a manuscript of any size. Similarly, any scene, irrespective of its popularity, has no chance of being in a manuscript of zero size. Thus, $\pi(0, n) = 0$ and $\pi(\theta, 0) = 0$.

An example of a function satisfying the above conditions is given by

$$\pi(\theta, n) = 1 - \exp(-cn\theta), \tag{4}$$

where c is a parameter such that the identity $\sum_j \pi(\theta_j, n) = n$ is satisfied. Its numerical value can be obtained iteratively by applying the Newton-Raphson method. Another example satisfying the first-derivatives and boundary conditions is

$$\pi(\theta, n) = 1 - (1 + d\theta)^{-cn}$$

where c and d are two non-negative parameters. A generalization allowing for changes in the degree of popularity of scenes is given by

$$\pi_t = \pi(\theta_t, n)$$

where t represents time.

3 Probability distribution of scene frequencies

3.1 Constant scene probabilities - Poisson distribution

The probability distribution of scene frequencies specifies the probability of the number of occurrences of scenes in a corpus of random illustrated manuscripts. First, we consider the simple case of a single scene j in a sample of m random manuscripts of equal size n. In this case, the probability of the occurrence of the scene j is constant for all manuscripts. It is equal to $\pi_j = \pi(\theta_j, n)$, where θ_j is the popularity index of the scene. The number of occurrences of scene j in the m manuscripts thus follows a binomial distribution with parameters m and θ_j, which for large m and small π_j can be approximated by a Poisson distribution with parameter $\lambda_j = m\pi_j$. The probability of observing exactly x occurrences of the scene j assuming that at least one is observed is given by

$$f_x(j, n) = \frac{\lambda_j^{x-1} \exp(-\lambda_j)}{(x-1)!}, \qquad x = 1, 2, ... \tag{5}$$

The Poisson distribution can be extended to apply to groups of scenes with the same popularity. Thus, if θ is the common popularity index of the scenes in a group of scenes, the probability distribution of the number of occurrences of the scenes in a large sample of m manuscripts of equal size n is expressed by

$$f_x(\theta, n) = \frac{\lambda^{x-1} \exp(-\lambda)}{(x-1)!}, \qquad x = 1, 2, ... \tag{6}$$

where

$$\lambda = m\pi(\theta, n).$$

3.2 Varying manuscript sizes - Negative binomial distribution

Now we derive the probability distribution of the number of occurrences of scenes of equal popularity in a large sample of m manuscripts of varying sizes. For the specification of scene probabilities in section 4, the manuscript size n is approximately proportional to the ratio λ/θ. Thus assuming that n, or, more conveniently the ratio λ/θ, follows a gamma distribution, we obtain

$$f_x(\theta) = \int_0^\infty f_x(\theta, n) dG(\lambda/\theta) = \frac{\Gamma(a+x-1)}{\Gamma(a)\Gamma(x)} \cdot \frac{(\theta b)^{x-1}}{(1+\theta b)^{a+x-1}}, \qquad x = 1, 2, ... \tag{7}$$

where $dG(.)$ represents the density function of the gamma distribution with shape parameter a and scale parameter b. The result is a shifted negative binomial distribution with parameters a and θb. The shift is one unit and reflects the fact that the distribution concerns Shahnameh scenes with at least one illustration.

The negative binomial distribution can be used to test the statistical significance of popularity of two or more scenes on the basis of the observed scene frequencies. If a scene is illustrated in one of say 100 random manuscripts, but another is illustrated in all manuscripts, it may be concluded without further analysis that the second scene is more popular than the first. But, if the observed scene frequencies are closer to each other, say twenty for the first and thirty for the second, drawing a conclusion on differences in scene popularity would require a test of significance.

3.3 Varying scene popularity - Generalized Waring distribution

The ratio $\theta b/(1+\theta b)$ in the expression of the negative binomial distribution is a monotonic transformation of the scene popularity index θ defined earlier. The transformed variable $\nu = \theta b/(1+\theta b)$ varies between 0 and 1, as θ varies between 0 and ∞. Thus, for least popular scenes $\nu = 0$ and for most popular scenes $\nu = 1$.

Assuming that scene popularity, in its transformed version, follows a beta distribution with parameters α and β, we derive the probability distribution of the number of occurrences of scenes of different degrees of popularity in a large sample of manuscripts of varying sizes. This is obtained by integrating the probability distribution $f_x(\theta)$ over the beta distribution

of the scene popularity ν as follows

$$f_x = \int_0^1 f_x(\theta)\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\nu^{\alpha-1}(1-\nu)^{\beta-1}d\nu \tag{8}$$

$$= \frac{\Gamma(a+x-1)}{\Gamma(a)\Gamma(x)}\cdot\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+x-1)\Gamma(a+\beta)}{\Gamma(\alpha+\beta+a+x-1)}, \qquad x=1,2,...$$

Except for a shift of one unit ($x = 1, 2, ...$ vs $x = 0, 1, ...$), this is the expression of the Generalized Waring distribution (Johnson, Kotz and Kemp 1993, p. 242), with mean equal to $\mu = 1 + a\alpha/(\beta - 1)$, and median, approximately, med$= a\alpha/\beta$. The distribution is used in lexicometry (Herdan, 1964, and Muller, 1977) and accident theory (Irwin, 1968). In lexicometry, it is used for quantifying literary styles and generally considered as the best existing model for fitting vocabulary curves (Holmes, 1985). In accident theory, it is used to fit the number of accidents occurring to people of varying proneness to accidents and different exposure to external risks of accident. It is shown to result from different hypotheses on the underlying process of risk exposure (Xekalaki, 1983).

3.4 Special cases

3.4.1 Waring distribution

For $\alpha = 1$ and $\beta = (c - a)$, the Generalized Waring distribution reduces to the Waring distribution with density function usually expressed in factorial terms

$$f_x = (c-a)\cdot\frac{(a+x-2)!}{(c+x-1)!}\cdot\frac{(c-1)!}{(a-1)!}, \qquad x=1,2,... \tag{9}$$

In the context of illustrated manuscripts, the Waring distribution corresponds to the assumption that the distribution of the popularity of scenes is skewed to the left, with relatively fewer popular scenes in relation to other types of scenes.

3.4.2 Yule distribution

It is obtained as a special case of the Waring distribution for $a = 1$, and $\beta = (c - 1)$. The density function is given by

$$f_x = \beta(\beta!)\cdot\frac{(x-1)!}{(x+\beta)!}, \qquad x=1,2,... \tag{10}$$

The Yule distribution corresponds to the assumption that the underlying distribution of sizes of manuscripts is exponential, while the scenes popularity follows a skewed beta distribution as in the case of the Waring distribution.

3.4.3 Diffused Generalized Waring distribution

For $\beta = 1$, the Generalized Waring distribution has infinite mean, and the resulting distribution is called here the Diffused Generalized Waring distribution. Its probability density function is given by

$$f_x = a\alpha \frac{\Gamma(a+x-1)\Gamma(\alpha+x-1)}{\Gamma(x)\Gamma(\alpha+a+x)}, \qquad x = 1, 2, ... \tag{11}$$

3.4.4 Diffused Waring distribution

A particularly simple distribution is obtained by setting $\alpha = 1$ in the above expression. The resulting density function is given by

$$f_x = \frac{a}{(a+x)(a+x-1)}, \qquad x = 1, 2, ... \tag{12}$$

The distribution has an infinite mean, and corresponds to a uniform distribution of ν. The probability function is given by $F_x = x/(a+x)$, $x = 1, 2, ...$, where a is the median. The continuous version of this distribution has the same form as the distribution of the popularity index θ. Thus, for $\alpha = \beta = 1$ and $a = 1/b$, the scene popularity (θ) and the scene frequency (x) have identical distributions, F, one in continuous form, the other discrete. In the Waring family, the Diffused Waring is the only distribution with this invariant characteristic.

4 Fitting the probability distributions to Shahnameh data

4.1 Mean-median estimation

A number of methods have been proposed for estimation of the parameters of the Generalized Waring distribution and its variants: factorial moments, mean-and-first frequency, maximum likelihood, among others. Here, we propose still another method applicable to situations where the mean may not exist. We call it "mean-and-median" estimation, as it consists of equating the theoretical mean and median of the distribution to the corresponding empirical values. When the theoretical mean does not exist, only the median is used. The results can be used as initial values for maximum likelihood estimation discussed in the next section.

In its generalized form, the Waring distribution has three parameters, a, α, and β. The first two are interchangeable and can be considered to have, a priori, equal values, $a = \alpha$. This observation limits the estimation procedure, at the first instance, to a or α, and b. The mean-median estimates for the Shahnameh data are summarized in Table 4.

Distribution	Mean-median estimation	a	α	β
Generalized Waring	$\beta = (\mu - 1)/(\mu\text{−med}-1)$	1.802	1.802	1.608
	$\alpha = a = \sqrt{(\beta \cdot \text{med})}$			
Waring	$\beta = (\mu - 1)/(\mu\text{−med}-1)$	3.247	1	1.602
	$a = \sqrt{(\beta \cdot \text{med})}$			
Yule	$\beta = (\mu - 1)/(\mu\text{−med}-1)$	1	1	1.602
Diffused Generalized	$\alpha = a = \sqrt{\text{med}}$	1.423	1.423	1
Diffused Waring	$a = \text{med}$	2.026	1	1

Table 4: Mean-median estimates for the Shahnameh data where $\mu = 6.390698$ and $\text{med} = 2.026316$ are, respectively, the empirical values of the mean and median of the distribution of the scenes frequencies.

4.2 Maximum likelihood estimation

The log-likelihood of a sample of M illustrated scenes from the Generalized Waring distribution, with m_x scenes appearing exactly x times, $x = 1, 2, ...$ is given by

$$L(a, \alpha, \beta) = \sum_x m_x \log(f_x), \tag{13}$$

where f_x is expressed by 8. Taking the mean-median estimates of a, α and β as initial values, the corresponding maximum likelihood estimates may be derived by maximizing $L(a, \alpha, \beta)$ using standard iterative non-linear multivariate local optimization programmes, function *nlminb*, S-PLUS for Windows (version 3.1). This programme was used for computing the results presented in Table 5, where the maximum likelihood estimates of the parameters of the Generalized Waring distribution and its four special cases have been computed for the Shahnameh data. A maximum of 9 iterations were necessary to reach convergence.

Distribution	Log-likelihood	a	α	β
Generalized Waring	$L = -1668.428$	1.667	1.667	1.625
Waring	$L = -1666.927$	3.390	1	1.472
Yule	$L = -1704.334$	1	1	0.739
Diffused Generalized	$L = -1674.107$	1.413	1.413	1
Diffused Waring	$L = -1674.556$	2.041	1	1

Table 5: Maximum likelihood estimates for the Shahnameh data.

According to these results, the largest log-likelihood is given by the two-parameter Waring distribution, followed by the Generalized Waring distribution. In the latter case, however, the result is critically dependent on the initial condition of having set $a = \alpha$. A slight departure from the

identity, say, $a = 1.800$ and $\alpha = 1.802$, everything else equal, led after 38 iterations to the estimators, $a = 11.854$, $\alpha = 0.584$ and $\beta = 2.208$, and the log-likelihood, $L = -1664.021$[3]. A similar change in the initial condition for the Diffused Generalized Waring distribution, however, did not altered the values of the maximum likelihood estimates, even when the initial values of a and α were set very far apart, $a = 10$, $\alpha = 1$.

The lowest log-likelihood is reached by the one-parameter Yule distribution, which gives also the widest difference between the maximum likelihood estimate and the initial mean-median estimate.

4.3 Goodness-of-fit test

The Generalized Waring distribution and its four special variants have been fitted to the range of scene frequencies of the Shahnameh illustrations in the Michigan database using the maximum likelihood estimates obtained in the preceding section. The results are shown in Table 6 below. The goodness-of-fit is measured by the chi-square value between the expected and actual number of scenes for each scene frequency, $x = 1, 2, ..., \max = 74$. Thus, the chi-square distribution has $74 - 1 = 73$ degrees of freedom, and its values for the different fits are given on the last line of Table 6[4].

It can be observed that the two diffused distributions, Diffused GW and Diffused Waring, give the best fit, as compared to the other three finite-mean distribution, Generalized Waring, Waring and Yule. In both cases, the deviations as measured by the chi-square value, χ^2, are not significant at the 5% level of significance for 73 degrees of freedom. The critical chi-square value is 94.4. This means that the three other finite-mean distributions have significant deviations, and all three fail the chi-square goodness-of-fit test. The worst fit is attained by the Yule distribution.

Between the two infinite-mean distributions, the Diffused Waring gives the best fit, and has the most tractable distribution function, with a single parameter. It appears to be a suitable candidate for describing scene frequencies of Shahnameh illustrations. It implies that the observed differences in the range of scene frequencies of Shahnameh illustrations in the Michigan database can be generated by a simple underlying probability distribution of scene popularity, after allowance for variation in the size of manuscripts. This tentative conclusion may be used in explaining the process which makes one scene more popular than another, and will be fur-

[3]The symmetry of a and α implies that both sets of values $a = 11.854$, $\alpha = 0.584$ and $a = 0.584$ and $\alpha = 11.854$ are maximizing the log-likelihood function.

[4]The figures in parentheses in the last column of Table 6 are the fitted values after proportionnal adjustments for maximum frequency described in section 4.4.

Scene frequency	Number of scenes	Expected number of scenes				
		Gen. Waring	Waring	Yule	Diff. GW	Diff. GW
1	204	195.5	195.2	274.1	208.9	212.1 (216.0)
2	117	116.6	112.9	100.1	109.1	107.1 (109.1)
3	57	73.3	72.2	53.5	65.8	64.6 (65.8)
4	45	49.3	49.5	33.9	43.9	43.2 (44.0)
5	29	35.0	35.7	23.6	31.3	30.9 (31.5)
6	35	26.0	26.8	17.5	23.4	23.2 (23.7)
7	20	19.9	20.7	13.6	18.2	18.1 (18.4)
8	13	15.7	16.4	10.9	14.5	14.5 (14.8)
9	14	12.6	13.2	8.9	11.9	11.9 (12.1)
10-14	41	37.9	39.7	28.4	37.1	37.2 (37.8)
15.19	22	18.5	19.3	15.5	19.4	19.5 (19.9)
20-29	21	17.6	18.1	16.9	19.9	20.2 (20.5)
30-39	11	8.2	8.2	9.2	10.2	10.3 (10.5)
40-49	6	4.7	4.6	5.9	6.2	6.3 (6.4)
50+	10	14.1	12.5	32.9	25.3	25.8 (14.4)
Total	645	645	645	645	645	645 (645)
χ^2	-	108.2	108.5	152.7	91.2	90.9 (81.8)

Table 6: Fitting Waring Distributions to range of frequencies of Shahnameh scenes.

ther examined in the next section with data from other Persian manuscript illustrations.

4.4 Finite sample correction

The Generalized Waring distribution and its variants assign positive probabilities to all scene frequencies, $x = 1, 2, ...$ with no maximum limit. In practice, however, x is bounded by the number of illustrated manuscripts in the dataset, since a particular scene cannot occur more than once in any given manuscript. The relatively high deviations between the observed and fitted values for the two extreme parts of the distribution in Table 6 can be explained by the Poisson approximation to the binomial distribution of scene occurrences assumed in the derivation of the Generalized Waring distribution and its variants in section 3 of the paper. It is known that as x increases, the binomial probabilities are first smaller, then larger, and then again smaller than the corresponding Poisson probabilities (e.g., Feller, 1966, p. 161).

The fit can thus be improved by limiting the Generalized Waring distribution or its variants to a finite support and adjusting the probabilities in proportion of the difference between the binomial and Poisson probabilities

for each scene frequency within the support of the distribution, i.e, the total number of illustrated manuscripts in the database (111 in the case of the Michigan Shahnameh database).

As a first approximation which corrects mainly at the upper tail of the distribution, the probabilities are adjusted proportionally so that the total over the maximum frequency is equal to one. In the case of the Diffused Waring distribution, the adjustment amounts to multiplying the density function f_x by the constant multiplying factor $\gamma = (1 + a/m)$, where m is the number of manuscripts in the database. It can be noted that as m gets larger, the correcting factor tends to one. The results for the Shahnameh database with $m = 111$ are given in parentheses in Table 6. It can be noted that the observed and fitted values for the last category 50+ are much closer now, and the chi-square value of the goodness-of-fit has correspondingly dropped to the value of $\chi^2 = 81.8$, despite a larger number of degrees of freedom ($110 = 111 - 1$).

4.5 Data from other literary texts

The one-parameter Diffused Waring distribution is fitted to the scene frequencies of illustrations from three other literary works: Khamseh Nizami, Haft awrang Jami, and Kalilah wa Dimnah.

4.5.1 Khamseh Nizami

The Khamseh Nizami, or the Quintet of Nizami, is a collection of five long poems in masnavi form, like the Shahnameh, composed by the Persian poet Nizami of Ganja in the later part of the twelve century, i.e., about two hundred years after Firdowsi's Shahnameh. The first poem, Makhzan al-Asrar (The Treasury of Mysteries) is a ethico-philosophical discourse, while the other four are romances - Khosrow u Shirin, Leyli u Majnun, Haft Paykar, and Iskandarnama. Like the Shahnameh, the Khamseh Nizami has been vastly illustrated, and continues to be a prolific source of inspiration for modern miniature and oil paintings in Iran.

The dataset used here consists of a listing drawn by Stchoukine (1977) of 1486 illustrations[5] from 69 manuscripts of the Khamseh dating from the 15th and 16th centuries in the Topkapi Saray Museum in Istanbul. It distinguishes 178 scenes, with the most frequently illustrated appearing in 45 of the manuscripts: "Khosrow watching, Shirin bathing outdoor".

The range of scene frequencies of the illustrations is shown in the two first columns of Table 7. The third column gives the fitted values based

[5]Stouchkine (1977) refers to 1612 illustrations in the text of his study, but only 1486 are listed in the index of the publication, the basis of calculations done here.

on the Diffused Waring distribution with maximum likelihood estimate of the parameter $a = 2.602$. The chi-square value is 78.9, above the critical threshold, 61.2, corresponding to $44 = (45 - 1)$ degrees of freedom, gives the second best fit, after the Generalized Waring with $a = \alpha = 2.105$ and $\beta = 1.611$.

4.5.2 Haft awrang Jami

The Haft awrang, or Seven Thrones, comprises seven poems also in masnavi form by the Persian Sufi poet, Abdul-Rahman Jami, completed in the late fifteenth century, i.e, about four hundred years after Firdowsi's Shahnameh and three hundred years after Nizami's Khamseh which it greatly emulates. Three of the poems are romances, Yusuf u Zulaykha, Leyli u Majun, and Salman u Absad. Another three are ethico-philosophical discourses, Silsilat al-dhahab, Tuhfat al-ahrar, and Subhat al-abrar, and the seventh - Khiradnama-i Iskandari is a combination of the two genres.

As part of a study of one of the most outstanding illustrated manuscript of the Haft awrang, the sixteenth century manuscript prepared for the Safavid prince Sultan Ibrahim, Marianna Shreve Simpson (1997) listed some 724 illustrations of the Haft awrang, under 196 scene titles. The scene with the highest frequency (53 occurrences) is from Yusuf wa Zuleika and depicts the episode of the "Egyptian women overwhelmed by Yusuf's beauty."

The range of scene frequencies and their fitted values are given in the two middle columns of Table 7. The Diffused Waring distribution with maximum likelihood estimate $a = 1.246$ gave the best fit. The resulting chi-square, 34.9, is substantially below the critical value of 69.3, corresponding to 52=(53-1) degrees of freedom at the 5% level of significance.

4.5.3 Kalilah wa Dimnah

The Kalilah wa Dimnah is one of the most popular books in the ancient world, written in the form of moral fables in which animals speak artfully, but mockingly on man's frailty and wit in his relentless battle for survival. The text has been vastly illustrated for over a thousand years, illustrated copies of which have survived from the thirteenth century onwards. Ernst Grube (1991) provides an extensive listing of 1786 illustrations of 64 manuscripts of the Kalilah wa Dimnah or its later versions, Anvar-i suhayli, Iyar-i danish, and Humayunnama. Some 457 illustrated scenes are distinguished, with scene frequencies ranging from 1 to 24. The two mostly illustrated scenes are "The killing of the camel" (24 occurrences) from the story of the Lion and the Bull, and "The monkey rides across the lake on the tortoise" (also 24 occurrences) from the story of the Monkey and the

Tortoise.

The range of scene frequencies and the corresponding fitted values obtained by the Diffused Waring distribution with maximum likelihood estimate of the parameter $a = 1.556$ are given in the last two columns of Table 7. Like in the case of the Khamseh Nizami, the fit of the Diffused Waring distribution for the Kalilah wa Dimnah data is not very close, the chi-square value, 47.2, being above the critical threshold of 34.9, corresponding to $23 = (24 - 1)$ degrees of freedom at the 5% level of significance. The best fit is obtained with the Generalized Waring distribution with estimated parameters $a = 115.143$, $\alpha = 0.555$ and $\beta = 22.832$.

Scene frequency	Khamseh Nizami		Haft awrang Jami		Kalilah wa Dimnah	
	Number	Fitted	Number	Fitted	Number	Fitted
1	55	49.4	77	87.3	168	179.2
2	28	27.9	49	33.5	79	78.4
3	8	18.0	18	17.7	51	44.0
4	10	12.5	14	11.0	41	28.2
5	11	9.2	10	7.5	25	19.6
6	8	7.01	3	5.4	15	14.4
7	6	5.6	5	4.1	14	11.0
8	2	4.5	2	3.2	7	8.7
9	3	3.8	2	2.6	11	7.1
10-14	13	12.0	7	7.8	26	21.7
15-19	9	6.5	5	4.0	14	11.1
20+	24	21.5	4	11.9	6	33.6
Total	178	178	196	196	457	457
χ^2	-	78.9	-	34.9	-	47.2

Table 7: Fitting the Diffused Waring Distributions to scenes frequencies of 3 literary works [Source: Stchoukine (1977), Simpson (1997) and Grube (1991)].

4.6 Concluding remarks

To assess the general applicability of the Diffused Waring distribution as a model of scene frequency in Persian illustrated manuscripts, it is necessary to widen the scope of literary texts examined, scrutinize more carefully the reliability of the underlying datasets, and evaluate the results when restricting the fit to manuscripts of particular periods and provenances. Comparison should also be made with other families of distributions, e.g., the Zipf-Estoup distributions used in linguistics (Johnson, et al., 1993, pp. 465-471).

Table 8: Michigan Shanameh database: Illustrated scenes and frequencies of occurences [Source: Norgren and Davis (1969)][6].

S001	49	Kayumars enthroned
S002	2	Kayumars instructs son to destroy divs
S003	4	Div Khazarvan fights Siyamak
S004	5	Hushang kills Black Div
S005	1	Hushang discovers fire while planning to kill dragon with stone
S006	1	Tahmuras enthroned
S007	3	Tahmuras, Divs
S008	7	Jamshid enthroned
S009	6	Jamshid teaches crafts
S010	3	Enthroned Jamshid carried by divs
S011	1	Zahhak consults physicians
S012	7	Jamshid halved before Zahhak
S013	19	Zahhak enthroned
S014	2	Zahhak relates dream to Arnavaz
S015	2	Zahhak, priest
S016	4	Feridun questions mother
S017	2	Zahhak hears Kaveh complaint
S018	2	Kaveh carries standard
S019	3	Feridun fords river on way to fight Zahhak
S020	5	Feridun with Jamshid sisters
S021	18	Feridun defeat Zahhak
S022	6	Feridun binds Zahhak
S023	7	Zahhak brought before Feridun
S024	11	Feridun brings Zahhak to Mount Damavand
S025	13	Feridun enthroned
S026	2	Jandal, Feridun envoy, before King Yemen
S027	6	King Yemen receives Feridun sons
S028	6	Feridun tests sons
S029	3	Sons, wives before Feridun enthroned
S030	2	Iraj with brothers Tur, Salm just before murder
S031	42	Tur, Salm slay Iraj
S032	9	Iraj coffin brought to Feridun
S033	7	Manuchehr, Iranians defeat Tur, Salm army
S034	21	Manuchehr kills Tur
S035	1	Qaran army beseiges Alan castle
S036	14	Manuchehr kills Salm
S037	7	Manuchehr enthroned
S038	1	Zal birth
S039	5	Simurgh carries Zal to its nest
S040	1	Sam recognizes son Zal in Simurgh nest
S041	3	Sam asks Simurgh to return Zal
S042	18	Simurgh brings Zal to Sam
S043	1	Sam brings Zal from mountains
S044	4	Prince Zal presented to Manuchehr
S045	1	Sam approaches Manuchehr
S046	1	Zal enthroned
S047	2	Zal meets Mehrab of Kabul
S048	3	Zal shoots fowl
S049	5	Zal by Rudabeh castle
S050	32	Zal meets Rudabeh
S051	1	Sindokht learns of Zal, Rudabeh
S052	1	Sam before Manuchehr
S053	1	Rudabeh father hears of Zal
S054	1	Queen Sindokht gives audience to Sam
S055	1	Zal, Sindokht
S056	2	Manuchehr, Zal
S057	2	Archmages question Zal
S058	1	Zal before Manuchehr after Archmages questioning
S059	5	Zal shows off to Manuchehr
S060	1	Sindokht meets Sam
S061	10	Zal marries Rudabeh
S062	24	Rostam birth

[6]The illustration titles have been edited for simplification: All definite and indefinite articles "the", "a", and "an" have been removed, as have the pronouns "he" and "his" and the verb "is"; The conjunction "and" has everywhere been replaced by a comma; Where appropriate, the preposition "of" has been removed, either directly as in "King of Mazandaran" changed to "King Mazandaran", by reversing the order of the connecting words as in "The coffin of Iraj" transformed to "Coffin of Iraj". Also, the words "Demon", "Persian" and "India" have been changed to "Div", "Iranian" and "Hind", respectively. Otherwise, the illustration titles and the frequencies are reproduced as in the original source, without corrections for known misclassification, redundancy and incompleteness.

S063	2	Sam sees Rostam portrait
S064	5	Sam visits Zal, Rostam
S065	27	Rostam slays White Elephant
S066	1	Young Rostam rides horse
S067	2	Manuchehr appoints Nowzar, dies
S068	2	Nowzar enthroned
S069	1	Qobad discusses death before battling Barman
S070	4	Qobad fights Barman
S071	2	Qaran army versus that of Afrasiyab
S072	5	Third battle between Nowzar, Afrasiyab
S073	4	Nowzar captured in battle with Afrasiyab
S074	2	Nowzar captive before Afrasiyab
S075	1	Zal joins Mehrab in battling Turanians
S076	6	Zal kills Khazarvan in battle
S077	1	Zal shoots Kalbad
S078	2	Nowzar just before execution
S079	15	Afrasiyab executes Nowzar
S080	3	Zav enthroned
S081	1	Zav makes peace with Turks, fixes boundary
S082	20	Rostam catches Rakhsh
S083	4	Rostam enthroned with Kay Qobad on Mount Alburz
S084	3	Rostam slays Qolun
S085	2	Kay Qobad enthroned
S086	2	Iranians versus Turanians
S087	3	Combat Rostam, Afrasiyab
S088	32	Rostam catches Afrasiyab by belt, lifts him
S089	2	Rostam pursues fleeing Afrasiyab
S090	2	Afrasiyab before Pashang
S091	1	Kay Qobad addresses Iranians
S092	5	Kay Kaus enthroned
S093	3	Div monstrel tempts Kay Kaus to invade Mazandaran
S094	1	Zal counsels Kay Kaus
S095	3	White Div captures, blinds Kay Kaus, army
S096	31	Rostam first feat: Rakhsh kills lion
S097	1	Rostam reproaches Rakhsh for killing lion
S098	1	Rostam washes Rakhsh in spring
S099	2	Rostam second feat: Finds spring, kills onager
S100	23	Rostam thrid feat: Slays dragon
S101	12	Rostam fourth feat: Cleaves witch
S102	1	Rostam fights watchman
S103	8	Rostam fifth feat: Lassoes Awlad
S104	13	Rostam sixth feat: Kills Arzhang
S105	1	Rostam confronts divs after killing Arzhang
S106	1	Rostam enters cave where Kay Kaus, men chained
S107	1	Owlad leads Rostam to White Div cave
S108	74	Rostam seventh feat: Slays White Div
S109	1	Rostam anoints Kay Kaus eyes White Div blood
S110	3	Rostam presents Kay Kaus letter to King Mazandaran
S111	1	Rostam charges Mazandaran army waving uprooted tree
S112	1	Rostam shakes hands with Kalahur Mazandaran hero
S113	3	Rostam battles champion Juya
S114	2	Kay Kaus paladins fight King Mazandaran
S115	4	Rostam overpowers King Mazandaran
S116	7	Battling Rostam, King Mazandaran changes into rock
S117	2	Rostam throws King Mazandaran as rock front Kay Kaus
S118	2	King Mazandaran dies
S119	1	Kay Kaus marries Sudabeh
S120	3	Kay Kaus captive before King Hamavaran
S121	1	King Hamavaran confines Kay Kaus, men in mountain
S122	1	Rostam lassoes King Sham
S123	2	Rostam lifts King Sham
S124	2	Rostam parleys with King Hamavaran
S125	1	Kay Kaus enthroned, orders world
S126	2	Div tempts Kay Kaus
S127	45	Kay Kaus air-borne
S128	1	Gudarz chastises Kay Kaus
S129	1	Rostam in Afrasiyab preserves
S130	1	Rostam encounters Afrasiyab across river
S131	2	Rostam battles Afrasiyab
S132	3	Iranians versus Turanians
S133	8	Rostam kills Alkus in battle
S134	2	Rostam tries to lasso Afrasiyab
S135	1	Rostam sleeps after hunt
S136	2	Turkmen capture Rakhsh
S137	18	Tamineh visits Rostam
S138	1	Sohrab chooses horse
S139	2	Sohrab captures Hajir
S140	4	Sohrab fights Gordafarid
S141	3	Rostam before Kay Kaus
S142	1	Rostam knocks down Tus in Kay Kaus presence
S143	1	Zhandeh Razm dead from blow of Rostam fist
S144	1	Sohrab, Hajir view Iranian camp
S145	3	Rostam meets Sohrab before combat
S146	2	First combat Rostam, Sohrab
S147	1	Rostam, Sohrab part after their first combat

S149	67	Last battle between Rostam, Sohrab
S150	17	Rostam mourns Sohrab
S151	1	Sohrab body carried to Zabolestan
S152	1	Tus, Gudarz encounter maiden in forest
S153	1	Feast for Siyavosh birth
S154	11	Siyavosh in Sudabeh dwellings
S155	1	Sudabeh accuses Siyavosh
S156	1	Sudabeh complains again to Kay Kaus
S157	67	Fire ordeal Siyavosh
S158	1	Siyavosh father embraces him after fire ordeal
S159	1	Siyavosh defeats Garsivaz
S160	1	Garsivaz before Siyavosh
S161	3	Siyavosh sends message to Afrasiyab
S162	1	Siyavosh welcomes Piran
S163	4	Afrasiyab greets Siyavosh
S164	27	Siyavosh plays polo before Afrasiyab
S165	2	Siyavosh displays archery skill before Afrasiyab
S166	6	Siyavosh hunts onager with Afrasiyab
S167	7	Siyavosh marries Farangis
S168	1	Siyavosh enters city gates
S169	1	Piran visits Siyavoshgerd
S170	1	Siyavosh hears of son Ferud birth
S171	1	Siyavosh, Garsivaz play polo
S172	1	Siyavosh on polo ground lifts mail garment
S173	3	Siyavosh battles Goruy
S174	1	Garsivaz slanders Siyavosh before Afrasiyab
S175	1	Garsivaz with Siyavosh
S176	4	Siyavosh taken prisoner
S177	58	Goruy executes Siyavosh
S178	3	Piran, Kay Khosrow before Afrasiyab
S179	1	Siyavosh mourned
S180	2	Sudabeh dies
S181	5	Faramarz slays Varazad
S182	3	Faramarz captures Sorkheh, Afrasiyab son
S183	5	Sorkheh executed
S184	1	Pilsam spears Giv
S185	20	Rostam slays Pilsam
S186	11	Rostam fights Afrasiyab, Human, Turanians
S187	2	Rostam on Turan throne
S188	10	Giv finds Kay Khosrow
S189	2	Kay Khosrow, Farangis watch Giv defeat Turanians
S190	4	Giv captures Piran
S191	2	Giv leads captive Piran before Kay Khosrow, Farangis
S192	25	Kay Khosrow crosses Oxus with Farangis, Giv
S193	1	Boatman asks pardon of Kay Khosrow for refusing passage
S194	6	Kay Kaus receives Kay Khosrow
S195	1	Kay Kaus writes patent for Giv
S196	2	Kay Khosrow army beseiges Bahman castle
S197	2	Kay Kaus abdicates in favor of Kay Khosrow
S198	12	Kay Khosrow enthroned
S199	9	Rostam, Zal before Kay Khosrow
S200	1	Kay Khosrow pledges vengence against Afrasiyab
S201	5	Kay Khosrow reviews troops
S202	2	Ferud views enemy from mountain top
S203	1	Ferud, Bahman parley
S204	1	Ferud battles Iranians
S205	5	Ferud slays Rivniz
S206	6	Ferud kills Zarasp
S207	2	Ferud battles Giv
S208	7	Bizhan battles Ferud
S209	1	Bizhan chases retreating Ferud
S210	2	Ferud retreats to fortress
S211	6	Ferud dead on throne; Jarireh, slaves tear their hair
S212	6	Jarireh commits suicide over her son body
S213	1	Tus army pillages Ferud castle
S214	1	Iranians come to mourn Ferud
S215	2	Bizhan after Palashan slaying
S216	1	Iranians versus Turanians
S217	3	Giv strikes crown from Tazhav head
S218	1	Bizhan pursues Tazhav who carries Espanay on horse
S219	7	Piran attachs Iranian camp at night
S220	1	Iranians withdraw after Piran defeats them
S221	1	Kay Khosrow hears of combat with Ferud
S222	1	Kay Khosrow disgraces Tus
S223	1	Representing Fariborz, Rohham seeks truce from Piran
S224	9	Turanians defeat Iranians
S225	3	Bizhan defends Kaveh standard
S226	2	Rivniz dies; men save crown
S227	6	Tazhav slays Bahram
S228	4	Giv captures Tazhav
S229	1	Iranians, Turanians battle second time
S230	2	Rostam asks Kay Khosrow to pardon Tus
S231	7	Tus battles Human
S232	7	Rohham subdues Bazur Turanian wizard
S233	4	Turanians use sorcery against Iranians in battle

S235	1	Iranians attack Turanians at night
S236	1	Rostam with Kay Khosrow
S237	3	Fariborz marries Farangis
S238	1	Iranians hold council at Mount Hamavan
S239	1	Piran consults Khagan Chin
S240	4	Kamus battles Giv, Tus
S241	2	Rostam comes to aid Iranians
S242	1	Piran meets Kamus
S243	1	Rohham fights Ashkabus
S244	59	Rostam slays Ashkabus, horse
S245	25	Rostam captures Kamus, kills him
S246	5	Rostam overturns Chengesh by seizing horse tail
S247	1	Rostam, Piran parley
S248	1	Rostam reproaches Piran
S249	2	Rostam battles force of Turan, Chin
S250	7	Rostam lifts Shangol from saddle but escapes
S251	3	Rostam battles Saveh
S252	1	Rostam kills Gahar
S253	54	Rostam pulls Khagan Chin from elephant by lasso
S254	2	Puladvand lifts Giv, Tus from their saddles
S255	1	Iranians call Rostam to battle
S256	27	Rostam vanquishes Puladvand
S257	1	Rostam remounts while Puladvand escapes
S258	42	Div Akwan flings Rostam into sea
S259	2	Rostam fights sea monster
S260	2	Rostam recovers Rakhsh
S261	1	Rostam attacks Afrasiyab herdsmen
S262	12	Rostam kills Div Akwan
S263	9	Kay Khosrow receives Rostam
S264	17	Bizhan hunts wild boar
S265	2	Manizeh nurse speaks with Bizhan
S266	9	Manizeh entertains Bizhan
S267	8	Bizhan taken prisoner, brought before Afrasiyab
S268	4	Piran stays execution of Bizhan
S269	2	Bizhan brought to pit
S270	2	Manizeh feeds Bizhan
S271	2	Rostam feasts with Giv
S272	9	Rostam dines with Kay Khosrow
S273	1	Rostam dressed as merchant before Piran
S274	3	Manizeh appeals to Rostam dressed as merchant
S275	57	Rostam rescues Bizhan from pit
S276	3	Rostam sacks Afrasiyab palace
S277	1	Rostam fights Afrasiyab
S278	3	Rostam, Bizhan battle Afrasiyab
S279	2	Kay Khosrow receives Rostam after Bizhan rescue
S280	1	Piran, Human parley
S281	1	Human challenges Gudarz
S282	1	Bizhan challenges Human
S283	44	Bizhan kills Human
S284	4	Bizhan slays Nastihan
S285	1	Gudarz receives Ruin
S286	1	Gudarz answers Piran letter
S287	2	Iranians versus Turanians
S288	1	Giv battles Piran
S289	4	Giv battles Lahhak, Farshidvard
S290	14	Fariborz slays Kalbad
S291	12	Giv battles Goruy
S292	4	Giv binds Goruy hands after defeating him
S293	9	Gorazeh kills Siyamak
S294	10	Feruhel pierces Zangoleh with arrow
S295	11	Barman battles Rohham
S296	12	Bizhan fights Ruin
S297	1	Bizhan leads away horse with Ruin body on it
S298	11	Hajir cleaves Sepahram helmet with sword
S299	8	Gorgin kills Andarman with arrow
S300	6	Gorgin leads away Andariman horse
S301	9	Zangeh, Akhast fight on horseback
S302	12	Barteh kills Kohram with sword
S303	2	Zangeh throws Akhast from horse
S304	15	Piran, Gudarz battle
S305	30	Piran escapes from Gudarz but later killed
S306	11	Gostaham kills Lahhak, Farshidvard
S307	2	Bizhan with wounded Gostaham
S308	2	Kay Khosrow mourns Piran, other Turanian dead
S309	3	Kay Khosrow kills Goruy
S310	1	Kay Khosrow receives Lahhak, Farshidvard bodies
S311	2	Kay Khosrow saves Gostaham
S312	2	Rostam, Kay Khosrow army ride to battle
S313	1	Afrasiyab sends Shideh to Kay Khosrow
S314	1	Rostam counsels Kay Khosrow on combat with Afrasiyab
S315	6	Kay Khosrow battles Shideh
S316	48	Kay Khosrow throws Shideh to ground
S317	11	Kay Khosrow army versus that of Afrasiyab
S318	2	Kay Khosrow encounters Ostogila, Ila, Borzoila
S319	2	Kay Khosrow, Iranians triumph

S321	14	Iranians capture Behesht, Afrasiyab fortress
S322	9	Afrasiyab family before Kay Khosrow
S323	2	Iranians repel Afrasiyab night attack
S324	1	Prisoners before Kay Kavus
S325	3	Kay Khosrow defeats, kills King Makran
S326	13	Kay Khosrow sees marvels of sea
S327	2	Kay Khosrow beseiges Gangdezh
S328	1	Kay Kaus, Kay Khosrow approach sacred fire
S329	8	Hum captures Afrasiyab
S330	4	Afrasiyab rises from water as Garsivaz tortured
S331	3	Kay Khosrow leads Afrasiyab by lasso
S332	2	Afrasiyab begs Kay Khosrow for mercy
S333	58	Kay Khosrow slays Afrasiyab while Garsivaz
S334	4	Zal, Rostam admonish Kay Khosrow
S335	16	Kay Khosrow makes appointments, gives up throne
S336	1	Kay Khosrow bid farewell to harem
S337	2	Kay Khosrow disappears
S338	8	Kay Khosrow paladins lost in snow, die
S339	1	Rostam, other paladins in mountain
S340	2	Lohrasp hears of Kay Khosrow disappearance
S341	30	Lohrasp enthroned
S342	2	Lohrasp speaks with son Goshtasp
S343	1	Goshtasp meets Zarir
S344	5	Goshtasp seeks work in Rum with smith Burab
S345	6	Goshtasp at Caesar Rum court
S346	24	Goshtasp slays wolf
S347	33	Goshtasp slays dragon
S348	3	Goshtasp plays polo before Caesar
S349	2	Goshtasp before Caesar
S350	6	Goshtasp slays Elys in Battle
S351	1	Zarir arrives at Caesar court
S352	1	Goshtasp welcomes Caesar
S353	1	Lohrasp receives Zarir, Goshtasp
S354	1	Lohrasp leaves wordly things
S355	6	Goshtasp enthroned
S356	8	Warriors of Gashtasp, Arjasp
S357	1	Nivzar dies in battle
S358	9	Biderafsh slays Zarir
S359	1	Esfandiyar father calls to him
S360	10	Esfandiyar kills Biderafsh
S361	15	Goshtasp puts son Esfandiyar in chains
S362	7	Arjasp horsemen kill Lohrasp
S363	1	Arjasp army sacks Balkh
S364	2	Arjasp flees from Goshtasp in combat
S365	2	Jamasp releases chained Esfandiyar
S366	2	Esfandiyar before Goshtasp
S367	2	Esfandiyar fights Turanians
S368	13	Esfandiyar lassoes Turanian Gorgsar
S369	2	Gorgsar captive before Esfandiyar
S370	1	Sultan Mahmud court
S371	19	Esfandiyar first exploit: Fights wolves
S372	24	Esfandiyar second exploit: Fights lions
S373	25	Esfandiyar third exploit: Battles dragon
S374	20	Esfandiyar fourth exploit: Slays sorceress
S375	31	Esfandiyar fifth exploit: Kills Simorgh
S376	7	Esfandiyar sixth exploit: Escapes from snow storm
S377	5	Esfandiyar seventh exploit: Crosses river, faces camel
S378	4	Esfandiyar kills Gorgsar during seventh exploit
S379	1	Esfandiyar disguised as merchant before Arjasp
S380	2	Esfandiyar sisters recognize him in merchant clothes
S381	2	Nush-azar slays Tarkhan
S382	23	Esfandiyar kills Arjasp, rescuses sisters
S383	1	Arjasp head flung from brazen hold
S384	4	Esfandiyar throws Kohram from saddle
S385	2	Kohram, Andariman executed
S386	1	Goshtasp receives Esfandiyar
S387	2	Esfandiyar demands Kingdom from father Goshtasp
S388	1	Zal greets Bahman
S389	26	Rostam kicks aside rocked pushed by Bahman
S390	1	Bahman delivers Esfandiyar message to Rostam
S391	6	Rostam, Esfandiyar parley
S392	6	Rostam feasts with Esfandiyar, Bahman
S393	3	Rostam boasts during trial of strength with Esfandiyar
S394	13	Combat Rostam, Esfandiyar
S395	3	Zavareh, Faramarz slay Nush-azar, Mehr-e-nush
S396	4	Wounded Esfandiyar, Rostam, Rakhsh seek shelter
S397	4	Simorgh aids wounded Rostam
S398	1	Esfandiyar, Bishutan discuss oncoming fight with Rostam
S399	3	Rostam, Esfandiyar fight again
S400	66	Rostam shoots double-pointed arrow in Esfandiyar eyes
S401	1	Rostam, Zal with dying Esfandiyar
S402	6	Esfandiyar mourned
S403	64	Rostam slays Shagad then dies
S404	1	Rostam funeral
S405	2	Faramarz mourns Rostam, Zawareh, father, brother

S407	1	Faramarz burns Shaghad body
S408	1	Girl takes snake from Rudabeh, mourning Rostam death
S409	1	Bahman enthroned
S410	16	Bahman kills Faramarz in combat
S411	1	Bishutan advises Bahman not avenge Rostam family
S412	1	Bahman releases Zal from bonds
S413	1	Humay puts her son in box in river
S414	1	Launderer wife rescues Darab from water
S415	4	Darab questions launderer wife about parentage
S416	1	Rashnawad reviews troops
S417	3	Rashnawad finds Darab asleep by castle ruins
S418	1	Darab, Rashnawad battle Rumis
S419	5	Darab receives crown
S420	1	Eskandar fed by mother
S421	2	Eskandar with Dara
S422	3	Eskandar army versus that of Dara
S423	36	Eskandar attends dying Dara
S424	6	Eskandar executes Dara murderers
S425	5	Eskandar enthroned
S426	2	Eskandar dictates letter to Delaray, Rushanak mother
S427	4	Eskandar meets Rushanak
S428	1	King Hind tells Mehran of dream
S429	1	Mehran foretells of Eskandar coming to King Hind
S430	1	Eskandar disctates letter to King Hind
S431	6	In India Eskandar receives daughter of King Hind
S432	2	Eskandar with smiths who made iron horse
S433	6	Eskandar army versus that of Fur
S434	15	Fur dies in battle against Eskandar
S435	13	Eskandar goes on pilgrimage to Ka'beh
S436	2	Queen Qaydafeh receives Eskandar portrait
S437	2	Eskandar leads army toward Andalos
S438	1	King Faryan before triumphant Eskandar
S439	14	Eskandar before Qaydafeh who holds portrait
S440	4	Eskandar with Queen Qaydafeh, her court
S441	1	Eskandar enthroned with Taynush
S442	4	Eskandar goes to land of Brahmans
S443	1	Serpents attack Eskandar
S444	3	Eskandar slays Habash monster
S445	12	Eskandar kills dragon
S446	18	Eskandar, Sage Khezr seek fountain of life
S447	3	Eskandar, talking birds
S448	4	Eskandar encounters Esrafil
S449	14	Eskandar builds wall against Gog, Magog
S450	9	Eskandar in land of talking trees
S451	1	Eskandar returns from China
S452	3	Approaching Babylon Eskandar meets Gushbestar
S453	3	At Babylon Eskandar sees baby, omen of death
S454	6	Eskandar dies
S455	4	Eskandar mourned
S456	1	Babak summons Sasan shepard, Ardashir father
S457	5	Ardawan disgraces Ardashir in dispute of hunt spoil
S458	1	Golnar sees Ardashir through Ardawan palace, falls in love
S459	2	Ardashir meets Golnar
S460	3	Golnar flees with Ardashir
S461	8	Ardashir, Ardawan battle
S462	4	Ardawan captive before Ardashir
S463	6	Ardawan executed
S464	1	Kurds defeat Ardashir
S465	7	Haftvad daughter story
S466	1	Ardashir views Haftvad castle
S467	9	Ardashir pours molten lead down worm of Haftvad throat
S468	5	Ardashir hangs Haftvad, eldest son Shahuy
S469	3	Ardashir enthroned
S470	1	Golnar tries to poison her husband Ardashir
S471	1	Ardashir minister stands before him
S472	11	Ardashir recognizes Shapur during polo game
S473	5	Mehrak daughter converses with Shapur
S474	2	Hormozd before Shapur during polo game
S475	1	Ardashir cedes throne to Shapur
S476	2	Shapur army versus that of Rum
S477	1	Bahram Bahramiyan enthroned
S478	5	Shapur enthroned
S479	5	Shapur besieges city of Arabian King Taer
S480	6	Taer beheaded before Shapur
S481	3	Shapur before King Rum
S482	3	Shapur captures King Rum
S483	8	King Rum before Shapur
S484	4	Shapur cuts off nose, ears of King Rum
S485	2	Shapur executes Mani
S486	2	Yazdgerd enthroned
S487	1	Yazdgerd offers robe of honor to Monzer Arab
S488	32	Bahram Gur hunts in company of Azadeh
S489	17	Bahram Gur mount tramples Azadeh
S490	7	Sea monster kills Yazdgerd
S491	2	Bahram Gur, Monzer, No'man receive Iranian delegation at Johram

S493	6	Bahram Gur enthroned
S494	1	Lambak, water carrier entertains Bahram Gur
S495	4	Bahram Gur rests in Baraham vestibule
S496	14	Drunken shoemaker rides on Bahram Gur escaped lion
S497	5	Bahram Gur, miller daughters
S498	1	Bahram Gur weds Nazyab
S499	2	Bahram Gur finds Jamshid treasures
S500	14	Bahram Gur kills dragon which devoured youth
S501	6	Bahram Gur at peasant house
S502	5	Bahram Gur with Barzin daughters
S503	3	Bahram Gur skewers doe, buck while hunting
S504	4	Bahram Gur listens to Arzu harp playing
S505	14	Bahram Gur hunts lions
S506	10	Bahram Gur hunts onager
S507	8	Bahram Gur soldiers versus those of Khagan Chin
S508	8	Bahram Gur takes prisoner Chinese Khagan
S509	1	Bahram Gur appoints Narsi, Khorasan viceroy
S510	1	Bahram Gur receives Caesar envoy
S511	4	Shangol entertains Bahram Gur
S512	13	Bahram Gur wrestles before Shangol
S513	9	Bahram Gur kills unicorn
S514	24	Bahram Gur kills dragon
S515	1	Bahram Gur kills monkey-lion before throne of Hind Shapur
S516	2	Bahram Gur chooses Sapinud for wife
S517	1	Bahram Gur, Khagan Chin exchange messages
S518	3	Bahram Gur travels with Sapinud
S519	3	Bahram Gur entertains Shangol, seven monarchs
S520	1	Bahram Gur bequeaths Kingdom to Yazdgerd
S521	7	Firuz, army fall into pit, die
S522	1	Balash enthroned
S523	3	Khoshnavaz flees from Sufaray in battle
S524	1	Khoshnavaz with Sufaray
S525	1	Iranians slander Sufaray before Qobad
S526	1	Shapur leads Sufary before Qobad
S527	1	Qobad enchained before Jamasp
S528	1	Razmehr helps Qobad escape
S529	2	Mazdak discusses philosophy before Qobad
S530	19	Nushirvan executes Mazdak, five followers
S531	11	Nushirvan enthroned
S532	4	Nushirvan displays prowess before Babak
S533	2	Nushirvan receives Gilan delegation
S534	2	Nushirvan consults Magi
S535	5	Nushirvan sacks fortress Arayesh-e Rum
S536	4	Nushirvan, Farfuryus fight
S537	1	Nushirvan receives tribute
S538	1	Nushirvan receives Mehras, Caesar envoy
S539	1	Nushirvan dictates letter to Ram Barzin
S540	6	Nushzad killed in battle with Ram Barzin
S541	5	Buzorjmehr interprets dream for Nushirvan
S542	6	Nushirvan finds male youth in harem
S543	2	Nushirvan executes youth, half-sister
S544	2	Nushirvan first banquet for Buzorjmehr
S545	2	Nushirvan fifth banquet for Buzorjmehr
S546	1	Nushirvan sixth banquet with Buzorjmehr, mobads
S547	1	Nushirvan seventh banquet for Buzorjmehr
S548	4	Two men poisoned by drink meant for Nushirvan
S549	3	Nushirvan hunting party after Mahbud murder
S550	1	Nushirvan questions Zarvan on role in Mahbud murder
S551	10	Nushirvan executes Zarvan, Jew
S552	6	Forces of Chin versus those of Haital
S553	2	Nushirvan dictates answer to Khagan letter
S554	5	Nushirvan envoy Mehran Setad in Khagan court
S555	12	Nushirvan receives Khagan daughter
S556	2	Buzorjmehr advises Nushirvan
S557	20	Nushirvan before chessboard sent by Hind Raja
S558	6	Buzorjmehr arranges chess pieces before Nushirvan
S559	4	Buzorjmehr invents Backgammon, shows how played
S560	1	Buzorjmehr shows Indian how to play Backgammon
S561	1	Gav, Talhand enthroned
S562	24	Gav, Talhand in combat
S563	3	Gav looks on Talhand body on battlefield
S564	2	Talhand mother learns of death
S565	1	Barzuy brings book "Kalileh Demneh" to Nushirvan
S566	1	Buzorjmehr disgraced
S567	1	Nushirvan receives Caesar locked chest
S568	2	Buzorjmehr divines contents of locked chest
S569	1	Caesar son receives Nushirvan letter
S570	6	Nushirvan defeats Rumis at their moat
S571	9	Caesar envoys before Nushirvan
S572	3	Hormozd enthroned
S573	1	Hormozd forces high priest to eat poisoned food
S574	1	Hormozd receives Nushirvan sealed casket
S575	1	Hormozd hunts
S576	1	Iranians under Kharrad defeat Khazars in Armenia
S577	2	Hormozd appoints Bahram Chubineh General

S579	1	Bahram Chubineh kills hay thief
S580	38	Bahram Chubineh defeats, kills Saweh
S581	2	Bahram Chubineh strikes Parmudeh
S582	1	Khagan Chin, Shah Hormozd parley
S583	11	Bahram Chubineh wears woman clothes from Hormozd
S584	1	Personage before palace
S585	1	Bahram Chubineh seated before young woman
S586	1	Gordyeh advices Bahram Chubineh
S587	2	Ayin Goshasp murdered
S588	1	Ayin Goshasp assassin executed
S589	2	Bendy, Gostaham blin Hormozd
S590	7	Khosrow Parviz enthroned
S591	4	Khosrow Parviz, Bahram Chubineh meet
S592	4	Khosrow Parviz, Bahram Chubineh dispute kingship
S593	20	Khosrow Parviz, Bahram Chubineh battle
S594	8	King Hormozd assassinated
S595	6	Bahram Chubineh enthroned
S596	2	Bahram Chubineh talks with Bahram, Siyavosh son
S597	7	Khossrow Parviz meets hermit
S598	4	Caesar of Rum receives Khosrow Parviz ambassadors
S599	1	Caesar consults astrologers
S600	11	Mystery of talisman
S601	1	Khorow Parviz approaches fire temple
S602	4	Khosrow Parviz with Maryam
S603	8	Khosrow Parviz, Bahram Chubineh fight second time
S604	5	Bahram Chubineh slays Kut Ruman
S605	6	Bahram Chubineh, Khosrow Parviz fight on mountain
S606	6	Angel Sorush rescues Khosrow Parviz
S607	5	Bahram Chubineh, Khosrow Parviz battle fourth time
S608	2	Bahram Chubineh wined by old woman
S609	1	Bahram Chubineh chases Nastuh
S610	4	Bahram Chubineh slays Maqatureh
S611	1	Lion-dragon devours Khagan daughter
S612	23	Bahram Chubineh kills lion-dragon
S613	15	Captured Qolun mortally wounds Bahram Chubineh
S614	1	Gordyeh speaks with Tovorg
S615	3	Gordyeh slays Tovora
S616	1	Khosrow Parviz executes Benduy
S617	1	Gordyeh kills Gostaham
S618	3	Gordyeh tells Khosrow Parviz of her achievements
S619	2	Khosrow Parviz praises Gordyeh
S620	1	Gordyeh presents equestrian cat to Khosrow Parviz
S621	17	Khosrow Parviz visits Shirin in her castle
S622	3	Khosrow Parviz hunts
S623	2	Shirin greets Khosrow Parviz returning from hunt
S624	2	Khosrow Parviz, Shirin in palace
S625	1	Khosrow Parviz celebrates Nowruz in garden
S626	18	Musician Barbad plays for Khosrow Parviz
S627	2	Shiruy followers find Khosrow Parviz in garden
S628	1	Zad Farrokh prepares to capture Khosrow Parviz
S629	3	Shiruy (Qobad) ascends throne
S630	2	Shiruy (Qobad) enthroned
S631	1	Khosrow Parviz answers son charges
S632	1	Mehr Hormozd offers to murder Khosrow Parviz
S633	19	Mehr Hormozd slays Khosrow Parviz
S634	3	Queen Shirin before Shiruy
S635	4	Shirin commits suicide near Khosrow Parviz coffin
S636	1	Shiruy dies
S637	1	Ardashir enthroned
S638	1	Puran-dokht enthroned
S639	1	Battle during Puran-dokht reign
S640	2	Yazdgerd enthroned
S641	19	Sa'd-e Vaqqas slays General Rustam
S642	2	Yazdgerd battles Mahuy
S643	2	Yazdgerd hides in Khosrow mill
S644	16	Miller assassinates Yazdgerd
S645	7	Bizhan kills Mahuy to avenge Yazdgerd

References

[1] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Volume I, Second edition. Wiley, New York.

[2] Grube, E. (1991). *Prolegomena for a Corpus Publications of Illustrated Kalilah wa Dimnah Manuscripts.* Islamic Art, New York, pp. 301-481.

[3] Herdan, G. (1964). *Quantitative Linguistics.* London, Butterworths.

[4] Holmes, D.I. (1985). The Analysis of Literary Style- A Review. *J. Roy. Statist. Soc. A* 148, 328-341.

[5] Irwin, J.O. (1968). The Generalized Waring Distribution Applied to Accident Theory. *J. Roy. Statist. Soc. A*, 205-225.

[6] Johnson, N.L., Kotz, S., and Kemp, A.W. (1993). *Univariate Discrete Distributions*, Second Edition. Wiley, New York, pp. 242-244, 271-272, and 278-279.

[7] Khalegi-Motlagh, D. (1995). *The Shahnameh*, Volumes I-IV. Mazda Publishers, Costa Mesa, California.

[8] Levy, R. (1990). *The Epic of the Kings.* Arkana, London.

[9] Mehran, F. (1996). The Use of Statistics in the Study of Persian Manuscript Paintings. *Proc. of the Third Iranian Statistics Conference*, Statistical Centre of Iran, Tehran.

[10] Mehran, F. (1997). Expected Number of Common Scenes in Illustrated Manuscripts. *Bul. of the Intern. Statist. Instit.*, 51st Session, Istanbul, Contributed Papers, Tome LVII, Book 2, pp. 207-208.

[11] Mehran, F. (1998). *Statistics and Persian Paintings: Atypical Illustrated Shahnameh Manuscripts.* Islamic Art, New York, forthcoming.

[12] Muller, C. (1977). *Principes et méthodes de statistique lexicale.* Hachette, Paris.

[13] Norgren, J. and Davis, E. (1969). *Preliminary Index of Shah-Nameh Illustrations.* The University of Michigan, Center for Near Eastern and North African Studies, Ann Arbor.

[14] Robinson, B.W. (1982). *Persian Painting and the National Epic.* Oxford University Press, Oxford.

[15] Safynedjad, J. (1995). Lithographic Shah-namehs (Shahnamehaye Chap Sanghi). In *Farsi* with summary English translation, Mirath-e Farhangi, Iranian Cultural Heritage Organization Periodical, No. 14, Tehran.

[16] Simpson, M. Shreve (1997). *Sultan's Ibrahim Mirza's Haft Awrang: a Princely Manuscript from Sixteenth-Century Iran.* Freer Gallery of Art, Smithsonian Institution, Washington, D.C., Appendice D, pp. 377-384.

[17] Stchoukine, I. (1977). *Les peintures des manuscrits de la "Khamseh" de Nizami au Topkapi Sarayi Mzesi d'Istanbul.* Institut Français d'archéologie de Beyrouth, Bibliothèque Archéologique et Historique - T. XCVI, Paris, Librairie Orientaliste Paul Geuthner, pp. 167-172.

[18] Warner, A.G. and Warner, E. (1925). *The Shahnama of Firdausi*, English translation. Kegan Paul, French, Trubner and Co., London.

[19] Xekalaki, E. (1983). The Univariate Generalized Waring Distribution in Relation to Accident Theory: Proneness, Spells or contagion?. *Biometrics* 39, 887-895.