Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

SAMPLING DISTRIBUTION OF THE VARIANCE

Pierre L. Douillet

Univ Lille Nord de France, F-59000 Lille, France
ENSAIT, GEMTEX, F-59100 Roubaix, France

ABSTRACT

Without confidence intervals, any simulation is worthless. These intervals are quite ever obtained from the so called "sampling variance". In this paper, some well-known results concerning the sampling distribution of the variance are recalled and completed by simulations and new results. The conclusion is that, except from normally distributed populations, this distribution is more difficult to catch than ordinary stated in application papers.

1 INTRODUCTION

Modeling is translating reality into formulas, thereafter acting on the formulas and finally translating the results back to reality. Obviously, the model has to be tractable in order to be useful. But too often, the extra hypotheses that are assumed to ensure tractability are held as rock-solid properties of the real world. It must be recalled that "everyday life" is not only made with "every day events" : rare events are rarely occurring, but they do.

For example, modeling a bell shaped histogram of experimental frequencies by a Gaussian *pdf* (probability density function) or a Fisher's *pdf* with four parameters is usual. Thereafter transforming this *pdf* into a *mgf* (moment generating function) by $mgf(z) = E_t(\exp zt)$ is a powerful tool to obtain (and prove) the properties of the modeling *pdf*. But this doesn't imply that a specific moment (e.g. μ_4) is effectively an accessible experimental reality.

This fact contains but is not limited to situations where these moments are infinite or undefined. For example, it is well known (Brown 2007) that the ratio of two standardized Gaussian variables is distributed according to a Cauchy *pdf*, so that the first moment exists only in principal value and the second moment is infinite. In fact, the mere difficulty occurs when these moments exists (this will be our hypothesis throughout the paper).

Moments of increasing index are increasingly dependent on the tails of the probability distribution, i.e are depending on increasingly rarer events and therefore are less and less accessible to experiment. Moreover, formulas that have to be used to evaluate these moments are increasingly complex and contain an increasing number of quite canceling terms, so that computation is unstable and propagates amplified uncertainties. This is even true for the simple "sample variance", that is our best guess of the "true" variance of the whole population.

The aim of this paper is to collect and illustrate some facts concerning this problem. The "Well Known Results" will be stated as such, while Theorem/Proposition will be reserved to new results or, at least, to results that are not usually emphasized. In Section 3, closed form results will be obtained for the very special situations when the sample size is either 2 or 3. It will be seen that even in this seemingly simple situation, general results are not easy to obtain.

In the remaining Sections, it will ever be assumed that samples contains at least four elements. Section 4 gives some experimental evidences, obtained using batches of $N = 200000$ independent samples. This value has been chosen in order to ensure "well shaped" curves... when such curves exist. It will be seen that these curves are often far away of the models generally used.

In Section 5, an algorithm is given that uses formal computing to re-obtain the formulas giving the best statistics for the moments of small index, and obtain these formulas and their Jacobian for $n = 11$ (new result). In Section 6 these formulas are used to determine the minimal size that a sample must have in order that a given statistic can be obtain from that sample. The paper ends with a concluding Section and some References.

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

2 NOTATIONS

Let us consider a probability set Ω and the associated r.v. (random variable) $\xi \in \Omega$. When relevant, the *pdf* of ξ is noted $\varphi(\xi)$. For a given $n \geq 2$, a sample ω of size n drawn "at random" from Ω will be an element of the set $\Phi \doteq \Omega^n$. By construction, sampling with replacement is assumed, ensuring that variables $x_i \in \omega$ are i.i.d.

Notation 2.1. The following equations summarize our notations :

$$\mu = E(\xi), \quad \mu_2 = \sigma^2 = var(\xi) = E\left((\xi-\mu)^2\right), \quad \mu_4 = E\left((\xi-\mu)^4\right) \tag{1}$$

$$m = E_\omega(x), \quad m_2 = s^2 = \frac{n}{n-1} var_\omega(x), \quad m_4 = \frac{n}{n-1} E_\omega\left((x-m)^4\right) \tag{2}$$

The expectations are noted by letter E. Without subscript, E denotes the Ω-expectation of a function of the random variable $\xi \in \Omega$. With subscript ω, E_ω denotes, for a given fixed sample ω, the ordinary mean value of a function of $x \in \omega$, so that $E_\omega(f(x)) = \sum_{x\in\omega} f(x)/n$. With subscript Φ, E_Φ denotes the Φ-expectation of a function of the sample ω, where the usual product measure is used over the set Φ.

The moments are noted by letters μ and m. Without subscript, μ denotes the expectation of variable $\xi \in \Omega$. With a subscript $i > 1$, μ_i denotes the corresponding *centered* moment. The symbol μ_1 will never be used. Letter m will be used in a similar manner to describe the mean and the corrected centered moments of variable $x \in \omega$ for a given sample $\omega \in \Phi$. Symbols σ, s will sometimes be used, when useful to avoid square roots.

When a formula doesn't contain μ, its proof is quite ever easier when assuming $\mu = 0$. This will be done without further mention.

Well Known Result 2.2. There are two usual measures for the skewness of a distribution. The Pearson's skewness is defined as $3(mean - median)/\sigma$ and ranges in $[-3..+3]$ while the Fisher's skewness, used throughout this paper and defined by :

$$\gamma_1 \doteq E\left((\xi-\mu)^3\right)/\sigma^3,$$

is not bounded. Common values are $\gamma_1(normal) = 0$, $\gamma_1(exponential) = 2$ and $\gamma_1\left(\chi^2_\nu\right) = \sqrt{8/\nu}$ where ν is the d.o.f. number.

Well Known Result 2.3. Let $A_0, A_1, \cdots, A_\nu$ be a partition of Ω such that $\forall j : p_j \doteq Pr(\xi \in A_j) > 0$. The $\chi^2_{Pearson}$ statistic of sample $\omega \in \Phi = \Omega^n$ is defined by :

$$\chi^2_{Pearson}(\omega) = \sum_{j=0}^{\nu} \frac{(n p_j - n_j)^2}{n p_j}$$

where n_j is the number of x_i that have fallen into A_j. Then, without any other assumptions, we have :

$$E_\Phi\left(\chi^2_{Pearson}(\omega)\right) = \nu, \quad var_\Phi\left(\chi^2_{Pearson}(\omega)\right) = 2\nu + \frac{1}{n}\left(3 - (\nu+2)^2 + \sum_0^\nu \frac{1}{p_j}\right)$$

giving a meaning to the standardized value $\chi^2_{std} = \left(\chi^2_{Pearson} - \nu\right)/\sqrt{2\nu}$ even when the $\chi^2_{Pearson}$ statistic is not χ^2_ν distributed.

3 RESULTS IN CLOSED FORM

Well Known Result 3.1. Considered from the Φ point of view, m and m_2 are random variables and we have :

$$E_\Phi(m) = \mu, \quad var_\Phi(m) = \frac{1}{n}\mu_2 \tag{3}$$

$$E_\Phi(m_2) = \mu_2, \quad var_\Phi(m_2) = \frac{1}{n}\left(\mu_4 - \mu_2^2 + \frac{2}{n-1}\mu_2^2\right) \tag{4}$$

Remark. Formula (4) is attributed to (Fisher 1929) by (Weatherburn 1962) and to ('Student' (Gosset, W.S.) 1908) by (Fisher 1929) himself. Many proofs can be given, among them Algorithm 5.2 on page 8.

3.1 Closed Forms when $n < 4$

Theorem 3.2. *Let φ be the p.d.f. of $\xi \in \Omega$. Then, for $n = 2, 3$, we have the following closed forms for the p.d.f. of m_2 :*

$$\begin{aligned} pdf_2(m_2) &= \sqrt{\frac{2}{m_2}} \int_{\mathbb{R}} \varphi(t)\, \varphi\left(t + \sqrt{2m_2}\right) \mathrm{d}t \\ pdf_3(m_2) &= \int_{t=0}^{t=s} \frac{4\sqrt{3}}{\sqrt{m_2 - t^2}} \int_{\mathbb{R}} \varphi(u-t)\, \varphi(u+t)\, \varphi\left(u + \sqrt{3m_2 - 3t^2}\right) \mathrm{d}u\, \mathrm{d}t \end{aligned}$$

Proof. Concerning $n = 2$, start from $1 = \iint \varphi(x)\, \varphi(y)\, \mathrm{d}x\, \mathrm{d}y$ and use $t = x$, $m_2 = \left(x^2 - 2xy + y^2\right)/2$ whose Jacobian is $J = y - x$. Chose the branch $y = t + \sqrt{2m_2}$, $J = \sqrt{2m_2}$ and compute $\iint \varphi(t)\, \varphi\left(t + \sqrt{2m_2}\right)/J\, \mathrm{d}t\, \mathrm{d}m_2$. Since both branches have equal contributions for a given m_2, $\iint = 1/2$ and pdf_2 follows. Concerning $n = 3$, start from $1 = \iiint \varphi(x)\, \varphi(y)\, \varphi(z)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z$ and use $t = (x-y)/2$, $u = (x+y)/2$, $m_2 = \left(x^2 - xy + y^2 - xz + z^2 - yz\right)/3$ whose Jacobian is $J = (2z - x - y)/3$. Chose the branch $z = u + \sqrt{3m_2 - 3t^2}$, $J = \sqrt{m_2 - t^2}/\sqrt{3}$ and compute $\iiint \varphi(u-t)\, \varphi(u+t)\, \varphi\left(t + \sqrt{3m_2 - 3t^2}\right)/J\, \mathrm{d}u\, \mathrm{d}t\, \mathrm{d}m_2$. Here again, a factor 2 appears to take both branches into account, and an extra factor 2 appears when using symmetry to restrict the integration domain to $x \geq y$ i.e. to $t \geq 0$. □

Remark. It can be checked that, applied to a normal variable, Theorem 3.2 leads to a χ^2 distribution (special cases of Well Known Result 3.4).

Theorem 3.3. *Let $\xi \in [-a, +a]$ be an uniform (continuous) random variable and $n = 3$ (the sample size). Then $s^2 = m_2 \in \left[0, 4a^2/3\right]$ with the following pdf :*

$$\begin{cases} pdf(m_2) = \frac{3\sqrt{3}}{a^2}\left(\frac{\pi}{6} - \frac{s}{2a}\right) & 0 < s < a \\ pdf(m_2) = \frac{3\sqrt{3}}{a^2}\left(\arcsin\frac{a}{s} - \frac{\pi}{3} - \frac{s}{2a} + \sqrt{\frac{s^2}{a^2} - 1}\right) & a < s \end{cases} \tag{5}$$

Proof. While integrating over $u \in \mathbb{R}$ in Theorem 3.2, the three factors product vanishes unless $u_1 \leq u \leq u_2$ where

$$\begin{aligned} u_1 &= \max(-a-t, -a+t, -a-W) = t - a \\ u_2 &= \min(a-t, a+t, a-W) = \min(a-t, a-W) \end{aligned}$$

and $\sqrt{3m_2 - 3t^2}$ has been shortened into W. In Figure 1, the discussion is drawn in the (m_2, t) plane. Zone A is characterized by $u_2 = a - W$ and zone B by $u_2 = a - t$, separated by the line $t = W$ i.e. $m_2 = 4t^2/3$. In order to enforce condition $u_1 \leq u_2$, zone B is bounded by $t = a$ and zone A by $m_2 = a^2 + (2t - a)^2/3$.

The inner integral evaluates to $3(2a - t - W)/\left(2W a^3\right)$ when $(m_2, t) \in A$, to $3(a-t)/\left(W a^3\right)$ when $(m_2, t) \in B$ and to 0 otherwise. Therefore, the outer integral has to be split into $t \in \left[0, s\sqrt{3}/2\right]$ and $t \in \left[s\sqrt{3}/2, s\right]$ when $s < a$ (the left dotted line) and split into $t \in [0, t_1]$, $t \in \left[t_2, s\sqrt{3}/2\right]$ and $t \in \left[s\sqrt{3}/2, a\right]$ when $s > a$ (the right dotted line). The rest of the computation is straightforward. It can be checked that (5) lead to $E(1) = 1$, $E_\Phi(m_2) = \mu_2 = a^2/3$ and $var_\Phi(m_2) = a^4/15$ as given by (4). □

Remark. The fact that $pdf(m_2)$ has a so complicated form, even for $n = 3$ and a so simple φ is another indication of the complexity of the question to solve.

3.2 Normal Distribution

Well Known Result 3.4 (Lukacs 1942)**.** Random variates m and m_2 are fully independent if and only if the sampled population Ω is normal. In such a case, $(n-1)m_2/\mu_2$ is χ^2_{n-1} distributed.

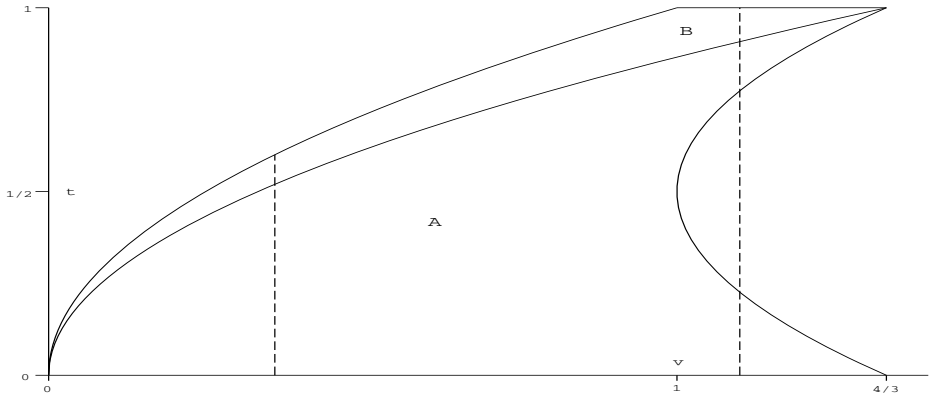


Figure 1: Graphical discussion of Theorem 3.3

Remark 3.5. Most of the time, Well Known Result 3.4 appears in the "Gaussian distribution" chapter of statistics books and is not recalled in the "χ^2" chapter. It should be emphasized that Gaussian distribution is not the paradigm but the exception when dealing with sample variance : the Gaussian distribution is the sole and only distribution such that sample mean and sample variance are independent. Therefore, the χ^2 model cannot even be applied as an approximate model for the sample variance relative to any non Gaussian distribution.

Remark 3.6. In the rest of the paper, non normal distributions of ξ will be considered. In order to facilitate comparisons between the induced distributions of the sample variance, it is of interest to compare their scaled squared coefficients of variation (sscv). From Well Known Result 3.4, the reference value of the sscv is :

$$sscv_{norm} = \frac{var_{\Phi}\left(s^2\right)}{\sigma^2} \times \frac{n-1}{\sigma^2} = \frac{var\left(\chi^2\right)}{E\left(\chi^2\right)} = 2 \tag{6}$$

4 EXPERIMENTING

This Section is devoted to some experimental results. To allow some comparisons, we start by a Gaussian example.

4.1 Normal Distribution

We have simulated $N = 200000$ samples from a Gaussian distribution, with sample size $n = 8$ and parameters $\mu = 0$, $\sigma = 7$. In Figure 2, we have plotted the experimental histogram of the sample variance (circles) together with the theoretical χ^2_7 (scaled) distribution (solid lines). The goodness of fit, as measured by $\chi^2_{Pearson} = 25.10$, i.e. $\chi^2_{std} = -1.28$ is excellent. A Gaussian curve, even with the required parameters, would not be the right model (dotted line) since $n = 8$ is far from infinity. Additionally, the experimental skewness of s^2 is $\gamma_1 \approx 1.07$, i.e. very close to the theoretical value, $\sqrt{8/7}$.

4.2 Uniform Distribution

When ξ is the discrete uniform distribution over the integer range $-a \leq \xi \leq +a$, the distribution of m_2 remains coarse, whatever the size N of the simulation. From $m_2 \leq n a^2/(n-1)$ together with $\left(n^2-n\right) m_2 \in \mathbb{Z}$, no more than $(an)^2$ different values of m_2 can occur. Moreover, this upper bound is not tight : when $a = 10$ and $n = 5$ the actual number of occurring values is 617, not 2500. The fact that not every integer is a square modulo $n^2 - n$ is one of the reasons of this drastic reduction. As a result, a batch involving $N = 200000$ samples leads to a very coarse distribution, as shown in Figure 3(a).

The corresponding histogram remains "rugged" as shown in Figure 3(b). Moreover, it appears that neither the normalized χ^2_{n-1} (dotted line) nor the adapted normal curve (solid line) provides even a rough approximation of the distribution.

Using a continuous uniform distribution leads to better looking experimental curves as shown in Figure 4 (here again, $a = 10$). But the departure from χ^2 remains in Figure 4(a) where $n = 5$ while a quite normal curve is obtained in Figure 4(b) where $n = 8$.

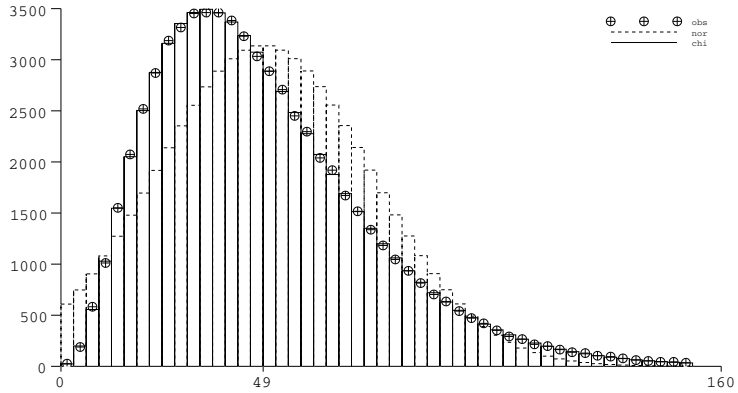
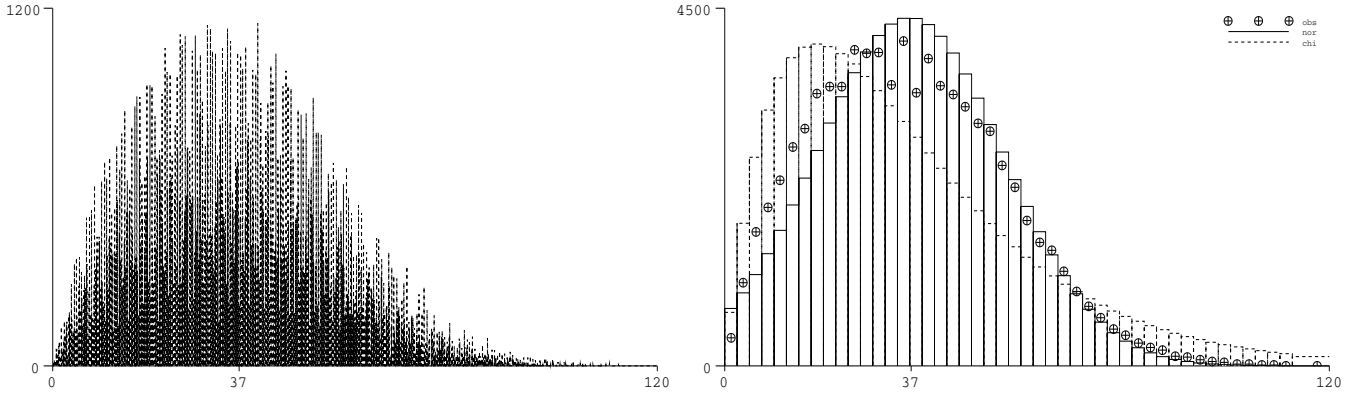
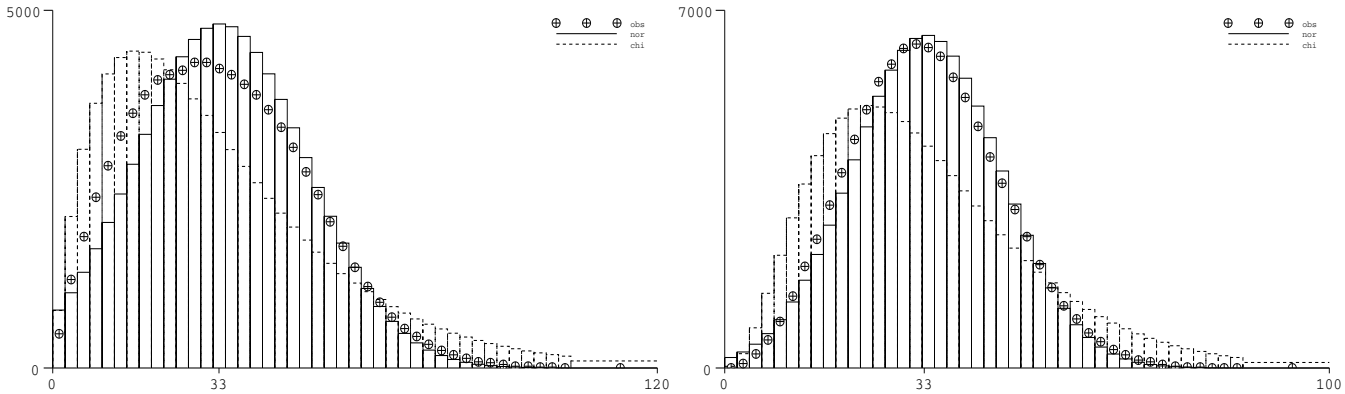


Figure 2: Normal law, $n = 8$



(a) showing coarsity (b) hidding coarsity

Figure 3: Discrete uniform distribution in $[-10, +10]$ (sample size $n = 5$)



(a) $n = 5$ (b) $n = 8$

Figure 4: Continuous uniform distribution in $[-10, +10]$

Proposition 4.1. *When ξ is a (continuous) uniform random variable in $[-a,+a]$ then $\mu_2 = a^2/3$, $\mu_4 = a^4/5$. The resulting scaled squared coefficient of variation –to be compared with* (6)*– is :*

$$sscv_{unif} = \frac{var_\Phi(m_2)}{\mu_2} \times \frac{n-1}{\mu_2} = \frac{4n+6}{5n} \tag{7}$$

Observed skewness are $\gamma_1 \approx 0.40$ for Figure 4(a) where $n=5$ and $\gamma_1 \approx 0.27$ for Figure 4(b) where $n=8$, far less than corresponding values for χ^2_{n-1} that are respectively $\gamma_1 \approx 1.41$ and $\gamma_1 \approx 1.07$. More results concerning skewness are given in Proposition 5.7.

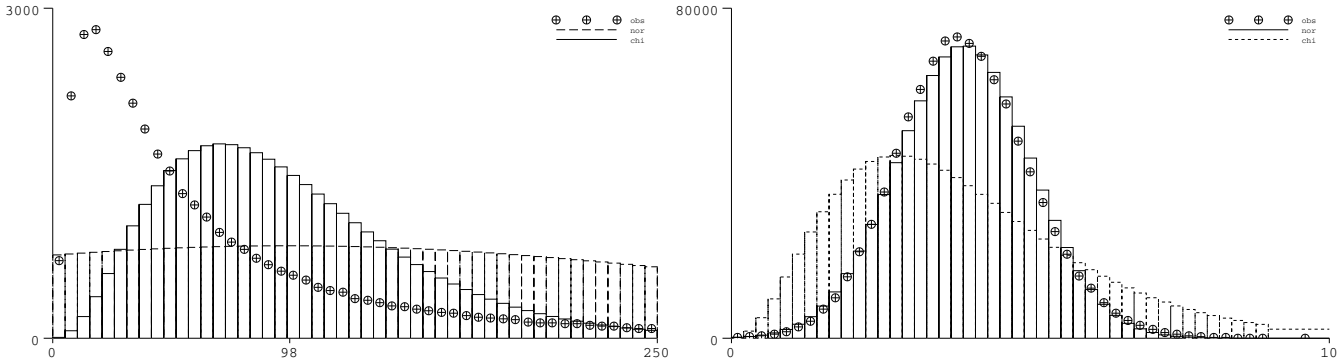
> Many statistics tend to be normally distributed as the data from which they are calculated are increased indefinitely; and this I suggest is the genuine reason for the importance which is universally attached to the normal curve (Fisher 1924).

4.3 Lognormal Distribution

Proposition 4.2. *When ξ is lognormal, let us define parameters M, K by $\ln M = E(\ln \xi)$ and $\ln K = var(\ln \xi)$. Then $\mu = M\sqrt{K}$, $\mu_2 = M^2 K(K-1)$. Moreover, $\mu_4 = M^4 K^2 (K-1)^2 \left(K^4+2K^3+3K^2-3\right)$. The resulting scaled squared coefficient of variation –to be compared with* (6)*– is :*

$$sscv_{logn} = \frac{var_\Phi(m_2)}{\mu_2} \times \frac{n-1}{\mu_2} = 2 + \frac{n-1}{n}(K-1)\left(K^3+3K^2+6K+6\right) \tag{8}$$

Figure 5(a) has been drawn with $M=7$, $K=2$ and $n=8$. Since ratio (8) is around 40, the observed skewness is huge ($\gamma_1 \approx 39$) leading to a curve that differs totally from either Gaussian or χ^2. On the contrary, as shown in Figure 5(b), a log scale gives a curve with $\gamma_1 \approx 0.05$ that fits really well with a Gaussian.



(a) flat (b) log

Figure 5: Lognormal distribution

4.4 Student's Like Distributions

Well Known Result 4.3 ('Student' (Gosset, W.S.) 1908)**.** For a sample drawn at random from a normal population, the statistic t defined by :

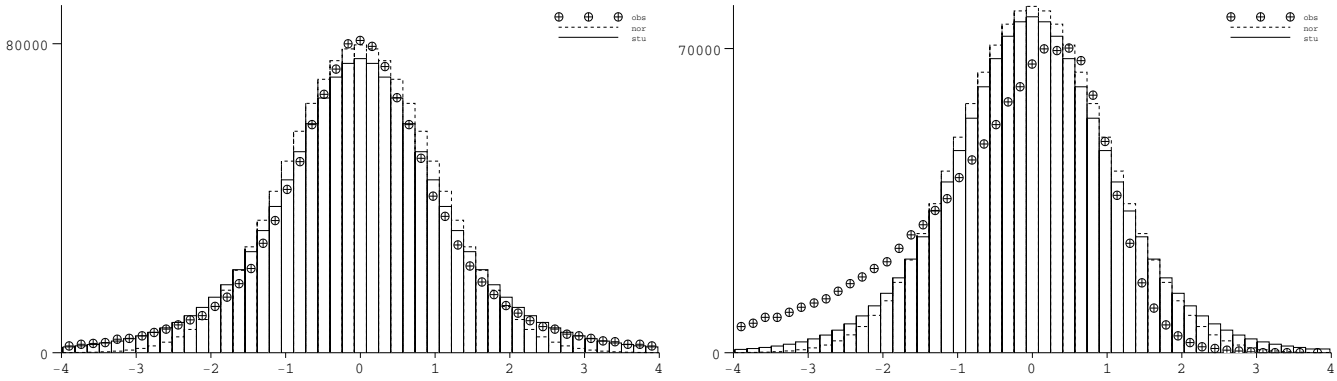
$$t = \frac{m-\mu}{s}$$

is distributed according to the Student-Fischer *pdf* :

$$\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$$

In order to see what happens when φ is not Gaussian, we have drawn the histograms of statistic t corresponding to Figure 4(a) (uniform, $n = 5$) and Figure 5(a) (lognormal, $n = 8$). In Figure 6(a), it can be seen that the tail of the experimental curve isn't very different from the corresponding Student-Fisher curve with $\nu = 4$ degrees of freedom. On the contrary, Figure 6(b) shows a very skew distribution, far different from the two tentative models.

In fact, the most surprising curve is the quite Gaussian curve associated with the uniform distribution. This can be related with the following fact. The intersection of hyperplane $m = constant$ and the hypercube Φ is an hyper-polygon. The more m is away from $\mu = 0$, the more this hyper-polygon shortens, leading to small values of s. Conversely, $E(s \mid m = 0)$ is as large as possible.



(a) From an uniform population

(b) From a lognormal population

Figure 6: Distribution of t-like statistics

5 ESTIMATING THE VARIATIONS OF THE SAMPLE VARIANCE

In many situations, estimating the variance of the population is not only a step for building confidence intervals around the mean, but is significant by itself. For example, an industrial product cannot be used for task requiring some precision when the adequate test shows a large variance. On the other hand, enforcing a useless precision will only induce additional costs. In production, a sudden increase in variability may indicate the appearance of a production fault (Crow, Davis, and Maxfield 1960).

In such situations, confidence intervals around the variance are to be discussed. When the moments of the population are not known *a priori*, formulas like (4) cannot be used, and must be replaced by formulas using the moments of the sample. Therefore, a method is needed to compute the expectation of these moments and of their products.

5.1 An Algorithm for the Expectation of Products

Definition 5.1. The m-degree of a monomial $\beta \doteq \prod m_k^{\beta_k}$ is $\deg_m \beta \doteq \sum \beta_k$ i.e. the number of factors m_k, distinct or not, occurring in β, while the x- degree of the same monomial is $\deg_x \beta \doteq \sum k\beta_k$ i.e. the number of factors x_i, distinct or not, occurring in the expansion of β into a polynomial in the x_i. Accordingly, $\deg_\mu \alpha \doteq \sum \alpha_j$ and $\deg_x \alpha \doteq \sum j\alpha_j$ are defined for a monomial $\alpha \doteq \prod \mu_j^{\alpha_j}$ depending on the population moments.

In order to obtain an unbiased statistic for a monomial α, we cannot simply substitute each μ_j by its unbiased statistic. This is obvious for terms like μ_j^2, but in fact this ever occurs since the moments of the sample are not (fully) independent when considered as random variates. We have to consider all the monomials β such that $\deg_x \beta = \deg_x \alpha$. After having computed each $E_\Phi(\beta)$, we can eliminate the irrelevant monomials in the μ_j in order to specifically isolate α.

Many research papers have been devoted to the computation of the $E_\Phi(\beta)$, (Tracy 1965) among them. These computations have been completely transformed by the formal computing tools that nowadays are largely accessible. To quote the Knuth's foreword to (Petkovsek, Wilf, and Zeilberger 1996) :

> Science is what we understand well enough to explain to a computer. Art is everything else we do. During the past several years an important part of mathematics has been transformed from an Art to a Science. No longer do we need to get a brilliant insight in order to evaluate sums of binomial coefficients, and

many similar formulas that arise frequently in practice. We can now follow a mechanical procedure and discover the answers quite systematically.

Algorithm 5.2. In order to obtain the closed form of a given $E_\Phi(\beta)$:

1) For each n in $[2,N]$ write β as an expanded polynomial in the x_i $(1 \le i \le n)$. Then substitute each power x_i^j $(j>1)$ by the corresponding moment μ_j and thereafter any remaining x_i by 0 ($\mu = 0$ is assumed). For each n, the result is a polynomial $P_n = \sum_\alpha c(n,\alpha) \times \alpha$ where the summation ranges over all the α such that $\mathrm{dg}_x \alpha = \mathrm{dg}_x \beta$ and the $c(n,\alpha)$ are rational numbers.

2) The closed form of each $c(n,\alpha)$ is a quotient of polynomials in n, whose degrees cannot exceed $\mathrm{dg}_x \beta$. They can be obtained by the algorithms described in (Petkovsek, Wilf, and Zeilberger 1996) and implemented as `gfun` in Maple (Salvy and Zimmermann 1994).

3) In fact, each denominator is a divisor of $n^p (n-1)^q$ where $p+q+2 = \mathrm{dg}_x \beta$ and $q+1 = \mathrm{dg}_m \beta$. Thereafter, it remains only to obtain the closed form of a polynomial in n from a list of values and Algorithm 5.3 can be used.

Proof. The rule concerning the degrees is obvious. The rationality of the $c(n,\alpha)$ comes from the binomial coefficients, and the specific value of the denominators in closed form comes from products of powers of $1/n$ (from the definition of the sample mean m) by powers of $1/(n-1)$ (from the definition of the m_k). □

Algorithm 5.3 (Newton)**.** In what follows ${}_k\delta_j$ denotes the j-th element of a list named ${}_k\delta$. Let ${}_0\delta$ contains the values taken by a polynomial p at some prescribed abscissas $n_0, n_1, \ n_N$. In other words, ${}_0\delta_j = p(n_j)$. In order to determine p, $N \ge \mathrm{dg}_p$ is assumed. For increasing k, compute the divided differences ${}_k\delta$ as defined by :

$$ {}_k\delta_j = \left({}_{k-1}\delta_{j+1} - {}_{k-1}\delta_j\right) / \left(x_{j+k} - x_j\right) $$

Then $p(x) = \sum_{k=0}^{N} {}_k\delta_0 \prod_{j=0}^{k-1} (x - n_j)$

Even automated, these computations are prone to errors and typos. For example, in page 208 of (Fisher 1929), we should have $\cdots 1120 * n - 1120)\mu_3^2 \mu_2$ in the $\mu\left(3^2 2\right)$ formula instead of $1120 * n + 1120$. While using his algorithm to compute cumulants by identifications, (Good 1977) has detected a typo in (ud Din 1954), while this article itself was signaling a typo from another author. An efficient test of correctness is the following :

Proposition 5.4. *For a given degree $d = \mathrm{dg}_x \beta$, the determinant of all the $E_\Phi(\beta)$ over the basis of all the α of the same degree splits into linear factors, namely n and $(n-1)$ in the denominator and $(n-j)$ where $2 \le j \le d-1$ in the numerator. For example, when $\mathrm{dg}_x \beta = 4, 5, 6, 10, 11$, we have :*

$$ \Delta_4 = \frac{(n-2)(n-3)}{n(n-1)}, \ \Delta_5 = \frac{(n-2)^2 (n-3)(n-4)}{n^3 (n-1)}, \ \Delta_6 = \frac{(n-2)^3 (n-3)^2 (n-4)(n-5)}{n^3 (n-1)^4} $$

$$ \Delta_{10} = \frac{(n-2)^{11} (n-3)^{10} (n-4)^7 (n-5)^5 (n-6)^3 (n-7)^2 (n-8)(n-9)}{n^{18} (n-1)^{22}} $$

$$ \Delta_{11} = \frac{(n-2)^{14} (n-3)^{12} (n-4)^{10} (n-5)^7 (n-6)^5 (n-7)^3 (n-8)^2 (n-9)(n-10)}{n^{28} (n-1)^{27}} \tag{9} $$

Proof. Clearly, the denominator splits into powers of n and $n-1$. To prove the form of the numerator, let us consider $1/\Delta$. When expressing a given α in the vector space spanned by the $E_\Phi(\beta)$, the l.c.m. of the involved denominators is necessarily $\prod(n-j)$ since (1) the degree of this polynomial has the maximal value and (2) each $n-j$ is required since the $E_\Phi(\beta)$ cannot form a basis when $j < \mathrm{dg}_x \alpha$. When computing the determinant of the α over the $E_\Phi(\beta)$, this denominator is elevated to power $\#\beta$: any exponent in the numerator cannot exceed this value (in fact, they form a decreasing sequence). □

Remark 5.5. It appears that no factors $n-1$ are canceling. When $\mathrm{dg}_x \beta = 11$, this leads to power :

$$27 = \sum_{\mathrm{dg}_x \beta = 11} (\mathrm{dg}_m \beta - 1)$$

At the same time, many powers of n are canceling, reducing the total degree of denominator from 126 (14 denominators, each of ninth degree) to only $27+28=55$.

5.2 Some Results

As a direct application of this algorithm, we have the following results.

Well Known Result 5.6. When $n>3$, let us define V by (10). Then $E_\Phi(V) = var_\Phi(m_2)$.

$$V \doteq \frac{1}{(n-2)(n-3)} \left(\sum (x-m)^4 - \frac{n^2-3}{n} m_2^2 \right) \tag{10}$$

Proof. The value of $var_\Phi(m_2)$ is given by (4). It's total degree is 4. It can be seen (Fisher 1929) that :

$$E_\Phi\left(m_2^2\right) = \frac{1}{n}\mu_4 + \frac{n^2-2n+3}{n(n-1)}\mu_2^2 \quad ; \quad E_\Phi(m_4) = \frac{n^2-3n+3}{n^2}\mu_4 + \frac{6n-9}{n^2}\mu_2^2 \tag{11}$$

The result follows by elimination. The $(n-2)(n-3)$ denominator "comes" from the fact that an undetermined expression is needed when $n=2$ and $n=3$ since, for these values, we have either $E_\Phi(m_4) = 2E_\Phi\left(m_2^2\right)$ or $E_\Phi(m_4) = E_\Phi\left(m_2^2\right)$, reducing the dimension of the vector space. □

Proposition 5.7. *The skewness of statistic m_2 is given by :*

$$\begin{aligned} \gamma_1(m_2) &= \frac{1}{\sqrt{n-1}} \frac{(n-1)^2\kappa_6 + 12n(n-1)\kappa_2\kappa_4 + 4n(n-2){\kappa_3}^2 + 8n^2{\kappa_2}^3}{\sqrt{n}\left((n-1)\kappa_4 + 2n{\kappa_2}^2\right)^{3/2}} \\ &= \frac{1}{\sqrt{n}} \left(\frac{\kappa_6 + 12\kappa_2\kappa_4 + 4{\kappa_3}^2 + 8{\kappa_2}^3}{\left(\kappa_4 + 2{\kappa_2}^2\right)^{3/2}} + \mathrm{O}\left(\frac{1}{n}\right) \right) \end{aligned} \tag{12}$$

Remark. When φ is Gaussian, all cumulants are 0 except from κ_2 and this formula reduces to $\sqrt{8/(n-1)}$, as it should be, since m_2 is χ^2_{n-1} distributed in this special case. In the general case, the distribution of m_2 is necessarily skew, and (12) shows that the asymptotic skewness is ever at most $\mathrm{O}(1/\sqrt{n})$.

Proposition 5.8. *The variance of estimator V is :*

$$\begin{aligned} var(nV) = {} & \frac{1}{n}\mu_8 - \frac{4(n-7)}{(n-1)n}\mu_2\mu_6 - \frac{\left(n^3-21n^2+47n-35\right)}{n(n-1)^3}{\mu_4}^2 + 4\frac{\left(2n^4-49n^3+230n^2-381n+210\right)}{n(n-1)^3(n-2)}{\mu_2}^2\mu_4 \\ & - 2\frac{(2n-3)\left(n^4-43n^3+285n^2-753n+630\right)}{n(n-1)^3(n-2)(n-3)}{\mu_2}^4 + 16\frac{\left(2n^4-16n^3+65n^2-115n+70\right)}{n(n-1)^3(n-2)}{\mu_3}^2\mu_2 - 8\frac{\left(n^2-4n+7\right)}{(n-1)^2n}\mu_3\mu_5 \end{aligned}$$

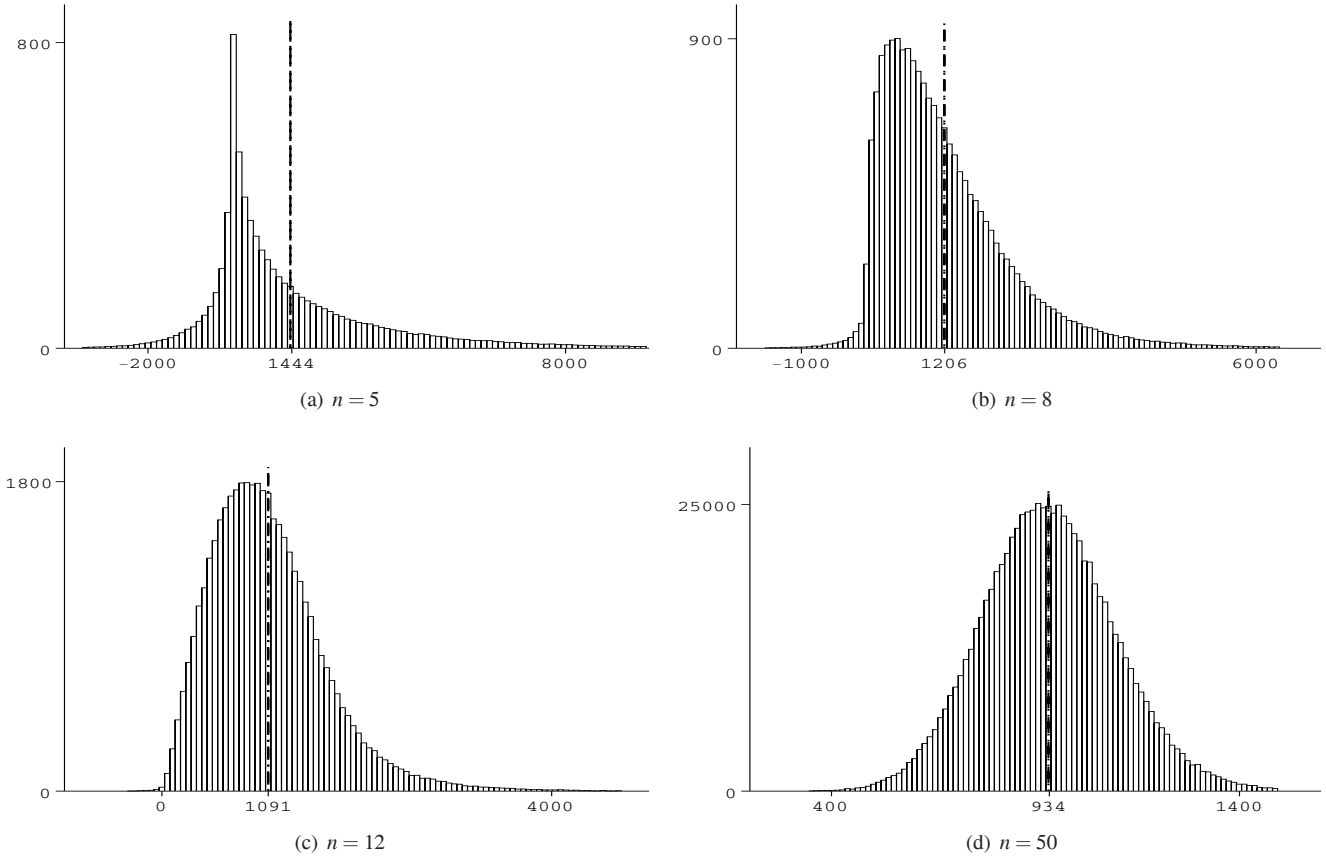
$$\begin{aligned} var(nV) = {} & \frac{1}{n}\kappa_8 + \frac{24}{n-1}\kappa_2\kappa_6 + \frac{2\left(17n^2-42n+29\right)}{(n-1)^3}{\kappa_4}^2 + \frac{8n(19n-37)}{(n-1)^2(n-2)}{\kappa_2}^2\kappa_4 \\ & + \frac{8n\left(7n^3-31n^2+33n+3\right)}{(n-1)^3(n-2)(n-3)}{\kappa_2}^4 + \frac{16n(3n-5)(4n-7)}{(n-1)^3(n-2)}{\kappa_3}^2\kappa_2 + \frac{16(3n-5)}{(n-1)^2}\kappa_3\kappa_5 \end{aligned}$$

Remark. According to (Fisher 1929), a better looking expression is obtained by transcoding moments into cumulants. But simplification is only on rational, exactly known coefficients. The underlying complexity, due to such a number of quite canceling terms remains the same : in the cumulants formula, all signs are positive, but the cumulants themselves aren't necessarily positive (even the cumulants of even index).

6 USEFUL AND USELESS STATISTICS

6.1 Uniform Distribution as an Example

When ξ is uniform over $[-10,+10]$, then $n \times var_{\Phi}(m_2) = 8000/9 + 20000/9/(n-1)$. An unbiased statistic for this quantity is nV where V is given in Well Known Result 5.6. In order to estimate the quality of this statistic, we have simulated four sets of $N = 200000$ samples, using respectively $n = 5, 8, 12, 50$, and plotted the results in Figure 7. In all cases the average of nV is as expected (the dashed line). But only the greatest value of n gives a nice shaped curve. For smaller values of n the distribution is really skew and for really small n, a noticeable part of the experimental values of V are negative (20% in Figure 7(a) where $n = 5$).



(a) $n = 5$ (b) $n = 8$

(c) $n = 12$ (d) $n = 50$

Figure 7: Experimental distribution of nV (when ξ is uniformly distributed)

6.2 Usefulness of a Statistic

Situation described in Subsection 6.1 shows that an "unbiased statistic" can be absolutely useless when dealing with small samples. In order to explore this question, we have to specify a border value beyond which noise will be considered as louder than signal.

Definition 6.1. A (positive) statistic α is *useless* (resp. *useful*) when its coefficient of variation is known to be greater (resp. *lower*) than $1/3$.

The idea beyond this definition has something to do with the notion of probable error. The *PE* is a deviation from the mean such that 50% of the population may be expected to lie between $\mu - PE$ and $\mu + PE$. This *PE* provides a rough perception of what happens, providing the following rule of thumb : below *PE*, don't discuss ; above *PE* begin to discuss.

In order to provide a similar criterion when the pdf is not easy to obtain, we have to select a threshold value for the coefficient of variation. Our choice of $1/3$ is based on the following reason. Probability distributions can be built such that quite all of the population lies inside of the "one sigma" range. But, outside the class room, these distributions are describing situation where rare events are a dominant feature, so that mean values have no more a clear factual meaning.

For the other situations, a great part of the population lies outside of the "one sigma" range. With our choice of factor $1/3$, this means that there is an important part of the population outside of $[2\overline{\alpha}/3, 4\overline{\alpha}/3]$ when the nominal value is $\overline{\alpha}$. Our feeling is that a not better known statistic should be discarded in any situation. Obviously, another choice of the factor, or a non symmetric interval (to take into account the unavoidable skewness of a positive variable), would be possible. But this would not change the mainlines of the argument.

Theorem 6.2. *The following statistics are, depending upon the distribution of $\xi \in \Omega$, useless when the sample size n is below the following values :*

pdf	m_2	*spec*	m_2^2	m_4	V	*spec*
uniform	*10*		*31*	*20*	*21*	
normal	*19*		*74*	*97*	*128*	
chi-square $\nu = 15$	*26*		*106*	*341*	*503*	
exponential	*72*	*36*	*305*	*1638*	*1934*	*148*

These values follow from the variances obtained in Subsection 5.2. Let us consider the Gaussian distribution. When $n = 19$, statistic $\alpha = 18 m_2/\mu_2$ is a χ^2_{18} random variable and $\overline{\alpha} \doteq E(\alpha) = 18$. The one sigma range of $\overline{\alpha}$ is $\left[18 \pm \sqrt{36}\right] = [12, 24]$ and therefore the diameter of the one sigma range for α is equal to $2\overline{\alpha}/3$. In this special case, the probability that α falls outside this range is easy to compute and amounts to 31%.

It has to be noticed that s^4 can be "useless" even if s^2 is "useful". This is partly related to $\mathrm{d}\left(x^2\right)/x^2 = 2\,\mathrm{d}x/x$ and partly related to the statistical nature of the involved quantities. Moreover, using explicitly that a variable is Gaussian results into $\mu_4 = 3\mu_2^2$ so that a useful statistic for $var_\Phi(m_2)$ as soon as $n > 74$ instead of $n > 128$ obtained by ignoring this relation.

For a chi-square distribution, a similar situation occurs. The only significant change is that border values are increasing.

6.3 Exponential Distribution

We will now examine in details what happens when $\xi \in \Omega$ is known to be exponentially distributed. This is a very strong hypothesis since it affirms that only one parameter is required to specify the population. If we are really sure of the validity of this hypothesis, we can lower the border of usability by an huge factor.

Statistic m is "useful" for estimating μ as soon as $n \geq 3$. Estimating μ_2 by m_2 will be foolish since a better statistic can be obtained via m^2. It can be seen that :

$$E_\Phi\left(m^{k+1}\right) = \frac{1}{\lambda^{k+1}} \frac{(n+k)!}{n!\, n^k}$$

so that $\alpha = m^2 n/(n+1)$ is an unbiased statistic for m_2, while $var(\alpha) = 2(2n+3)/n/(n+1)/\lambda^4$: α is a "useful" statistic for μ_2 when $n > 36$ (column spec in the table). The same argument holds for $var_\Phi(m_2)$ (second column spec).

Among $N = 40000$ samples of size $n = 36$ drawn at random from an exponential population, with $1/\lambda = 10$, the following values have been obtained. When using $\alpha = m$ as statistic for μ, then $542 + 1240$ values fall outside $[2\overline{\alpha}/3, 4\overline{\alpha}/3]$, i.e. a proportion of 4.5% (cf. $\sqrt{var(m)} = \mu/6$: it's a two-sigma interval for a variate not so far from normality). When using $\alpha = m^2$ as statistic for μ_2, then $5869 + 6104$ values fall outside of $[2\overline{\alpha}/3, 4\overline{\alpha}/3]$, i.e. a proportion of 30%, and 389 values outside $[\overline{\alpha} \pm \overline{\alpha}]$, i.e. around 1%. When using $\alpha = m_2$ as statistic for μ_2, then $9812 + 7692$ values fall outside of $[2\overline{\alpha}/3, 4\overline{\alpha}/3]$, i.e. a proportion of 44%, and 1492 values outside $[\overline{\alpha} \pm \overline{\alpha}]$, i.e. around 4%.

7 CONCLUSION

The χ^2 distribution is often used to model the behavior of the s^2 statistics. This is obviously valid for normal variates, but this doesn't apply to the general case. The first reason is that, apart from this special case of the normal distribution, statistic s^2 (the sample variance) is not independent from the sample mean m so that only a joint distribution can make sense when computing confidence intervals. Moreover, it appears that in some circumstances, s^2 is quite-normally distributed even for small values of n, while in some other circumstances s^2 is distributed in a very different way. Such a behavior must be taken into account when determining the minimal size of a sample for various statistic tests, aggravating the problematic described in (Chan, Hrobjartsson, Jorgensen, Gotzsche, and Altman 2008).

By the way, a method for computing and ***checking*** the product moments has been developed: the determinant of all the product moments of a given degree must split into a product of simple linear factors. The value of Δ_{11} – 9 on page 8– is a new result, while values of Δ_d were not explicitly stated in the many research papers devoted to the product moments of degree d less than 11.

Acknowledgement. We like to thank the Anonymous Referees for their helpful and constructive as well as very extensive comments on the submitted manuscript.

REFERENCES

Brown, K. 2007. Ratio populations. Available via http://mathpages.com/home/kmath042/kmath042.htm [accessed April 3, 2009].

Chan, A., A. Hrobjartsson, K. J. Jorgensen, P. C. Gotzsche, and D. G. Altman. 2008. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ* 337 (dec04-1): a2299–.

Crow, E. L., F. A. Davis, and M. W. Maxfield. 1960. *Statistics manual*. Unabridged reprint of NAVORD Report 3360–NOTS 948, U.S. Naval Ordnance Test Station ed. Dover Publications.

Fisher, R. A. 1924. On a distribution yielding the error functions of several well known statistics. In *Proc. Internat. Math. Cong. Toronto*, Volume 2, 805–813.

Fisher, R. A. 1929. Moments and product moments of sampling distributions. *Proc. London. Math. Soc.* 30:199–238.

Good, I. J. 1977. A new formula for k-statistics. *The Annals of Statistics* 5 (1): 224–228.

Lukacs, E. 1942, March. A characterization of the normal distribution. *The Annals of Mathematical Statistics* 13 (1): 91–93.

Petkovsek, M., H. S. Wilf, and D. Zeilberger. 1996. *A=b*. AK Peters, Ltd.

Salvy, B., and P. Zimmermann. 1994. Gfun: A maple package for the manipulation of generating and holonomic functions in one variable. *ACM Transactions on Mathematical Software* 20 (2): 163–177.

'Student' (Gosset, W.S.) 1908. On the probable error of a mean. *Biometrika* 6:1–25.

Tracy, D. S. 1965, September. *Finite moment formulae and products of generalized k-statistics with a generalization of Fisher's combinatorial method*. PhD in Mathematics, L.S.A. College, The University of Michigan, Ann Arbor, MI.

ud Din, Z. 1954. Expression of the k-statistics k9 and k10 in terms of power sums and sample moments. *Ann. Math. Statist* 25 (4): 800–803.

Weatherburn, C. 1946, 1962. *A first course in mathematical statistics*. 2 ed. London: Cambridge University Press.

AUTHOR BIOGRAPHY

PIERRE L. DOUILLET is an Associate Professor in Applicable Mathematics at the Ecole Nationale Superieure des Arts et Industries Textiles (Textile Engineers Institute), Roubaix, France. He studied Mathematics and Computer Science at the Ecole Normale Superieure (ENS Cachan, 1969-1974) and received his M.Sc and Ph.D. degrees in Computer Science from the Science University of Paris (Jussieu, Paris VI). His research interests include formal computing, approximations, network performances evaluation, operations research and computer's security. His email address is <douillet@ensait.fr> and his web page is <www.douillet.info>