A SIMPLE CHI-SQUARE TEST

OVE DITLEVSEN

1. Statistics based upon the distributions of extreme values. The distribution functions for the largest and smallest value, respectively, in a sample of m members drawn independently from a population with the distribution function $F(x)$ are

$$G_{\max}\left(x;\frac{1}{m}\right) = F(x)^m, \quad G_{\min}\left(x;\frac{1}{m}\right) = 1-[1-F(x)]^m ,$$

where m is a positive integer.

If, more generally, the distribution functions

$$G_{\max}(x;\alpha) = F(x)^{\frac{1}{\alpha}} \tag{1.1}$$

$$G_{\min}(x;\alpha) = 1-[1-F(x)]^{\frac{1}{\alpha}} , \tag{1.2}$$

where α is any positive number, are considered, a first exercise would be to calculate the maximum likelihood estimates of α, when a sample $x_1, x_2, \ldots, x_n$ of n independent members from the relevant populations is given. In the following, only distributions of the continuous type will be considered. For (1.1) the likelihood equation becomes

$$\begin{aligned}\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial \ln \left[\frac{1}{\alpha} F(x_i)^{\frac{1}{\alpha}-1} f(x_i)\right]}{\partial \alpha} &= \sum_{i=1}^{n} \frac{\partial}{\partial \alpha}\left[-\ln \alpha + \left(\frac{1}{\alpha}-1\right)\ln F(x_i)\right] \\ &= -\frac{n}{\alpha} - \frac{1}{\alpha^2}\sum_{i=1}^{n} \ln F(x_i) = 0 ,\end{aligned} \tag{1.3}$$

where $f(x) = dF(x)/dx$, assuming this derivative to exist and be continuous piece by piece. The unique solutions of (1.3), and of a similar equation corresponding to (1.2), are

$$\alpha^* = -\frac{1}{n}\sum_{i=1}^{n} \ln F(x_i) \tag{1.4}$$

$$\alpha^* = -\frac{1}{n}\sum_{i=1}^{n} \ln[1-F(x_i)] . \tag{1.5}$$

The next step is to find the distribution of these statistics.

2. The distribution of the estimates α^*. It is easy to obtain the distribution of α^* by use of characteristic functions [1, p. 185]. Corresponding to the variable $-\frac{1}{n}\ln F(\xi)$, the characteristic function is defined by

$$\varphi(t) = \mathcal{m}\left\{e^{it\left(-\frac{1}{n}\ln F(\xi)\right)}\right\}, \tag{2.1}$$

where ξ has the distribution function $F(x)^{\frac{1}{\alpha}}$, and where i is the imaginary unit. Using (1.1) and the substitution $u = F(x)$, the formula (2.1) becomes

$$\varphi(t) = \int_{-\infty}^{\infty} e^{-it\frac{1}{n}\ln F(x)} dF(x)^{\frac{1}{\alpha}} = \frac{1}{\alpha}\int_0^1 u^{-\frac{it}{n}+\frac{1}{\alpha}-1} du = \frac{n}{n-it\alpha}.$$

In virtue of the independence of the terms in (1.4), the characteristic function of α^* is then [1, p. 188]:

$$\varphi_1(t) = \varphi(t)^n = \left(1-\frac{it\alpha}{n}\right)^{-n}.$$

Introducing the variable $2n\frac{\alpha^*}{\alpha}$, the characteristic function will be

$$\varphi_2(t) = \mathcal{m}\left\{e^{it\,2n\frac{\alpha^*}{\alpha}}\right\} = \varphi_1\left(\frac{2nt}{\alpha}\right) = (1-2it)^{-\frac{2n}{2}},$$

which is the characteristic function corresponding to the χ^2-distribution with $2n$ degrees of freedom [1, p. 233].

This interesting and useful result, which, by a similar calculation, also appears to be valid for (1.5), can, with $1/\alpha = m$, be stated as follows:

The random variables

$$-2m\sum_{i=1}^{n}\ln F(x_i) \tag{2.2}$$

$$-2m\sum_{i=1}^{n}\ln[1-F(x_i)], \tag{2.3}$$

corresponding to samples from populations with the distribution functions $F(x)^m$ and $1-[1-F(x)]^m$, respectively, are distributed like χ^2 with $2n$ degrees of freedom, if $F(x)$ is of the continuous type.

Naturally this is only an elementary generalization of the fact that $\eta = F(\xi)$ is uniformly distributed on the interval $[0,1]$ when $P(\xi \leq x) =$

$F(x)$ (the probability integral transformation). It is then clear that $-2 \ln \eta$ has the frequency function 0 for $x < 0$ and

$$\frac{d}{dx} P(-2 \ln \eta \leqq x) = \frac{d}{dx} P(\eta \geqq e^{-\frac{1}{2}x}) = \frac{d}{dx}(1 - e^{-\frac{1}{2}x}) = \tfrac{1}{2} e^{-\frac{1}{2}x}$$

for $x \geqq 0$, which is the frequency function for χ^2 with 2 degrees of freedom.

For $m = 1$, the two populations become identical with the distribution function $F(x)$. This suggests that (2.2) and (2.3) may be used as test statistics for the hypothesis that the population has the distribution function $F(x)$. The disadvantages of the usual χ^2-test, which demands a grouping of the sample (see [2]), and which is only asymptotically correct, is not present in this test, although naturally its usefulness depends upon the power of the test. This question will be elucidated in a number of examples in the following.

For $m > 1$, there are also some cases where (2.2) and (2.3) may be used as test statistics. In a number of physical measurements, it is only the largest or smallest value among m values that is registrated. An obvious case is the measurement of the strength of a chain with m links. The strength of the weakest link determines the strength of the chain as a whole. A hypothesis about the distribution of the strength of a link selected at random may then be tested by the statistic (2.3).

3. The joint asymptotic distribution for the two statistics corresponding to $m = 1$. Let

$$\tau_1 = -2 \sum_{i=1}^{n} \ln F(x_i) \tag{3.1}$$

$$\tau_2 = -2 \sum_{i=1}^{n} \ln [1 - F(x_i)] . \tag{3.2}$$

The first and second order moments are

$$\mathcal{M}\{\tau_1\} = \mathcal{M}\{\tau_2\} = 2n, \quad \mathcal{V}\{\tau_1\} = \mathcal{V}\{\tau_2\} = 4n$$

$$\mathcal{M}\{\tau_1 \tau_2\} = 4 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{M}\{\ln F(x_i) \ln [1 - F(x_j)]\}$$

$$= 4n\mathcal{M}\{\ln F(\xi) \ln [1 - F(\xi)]\} + \mathcal{M}\{-2 \ln F(\xi)\} \mathcal{M}\{-2 \ln [1 - F(\xi)]\}(n^2 - n) ,$$

where [3, formula 440 p. 283]:

$$\mathcal{M}\{\ln F(\xi)\ln[1-F(\xi)]\} = \int_0^1 \ln u \ln(1-u)\,du$$

$$= \lim_{(c_1,c_2)\to(0,1)} [2u + u\ln u \ln(1-u) + (1-u)\ln(1-u) - u\ln u]_{c_1}^{c_2}$$

$$+ \int_0^1 \frac{\ln u}{1-u}\,du = 2 - \frac{\pi^2}{6},$$

such that the coefficient of correlation becomes

$$\varrho\{\tau_1,\tau_2\} = \frac{\mathcal{M}\{\tau_1\tau_2\} - \mathcal{M}\{\tau_1\}^2}{\mathcal{V}\{\tau_1\}} = 1 - \frac{\pi^2}{6} \approx -0.645\,.$$

From the *central limit theorem* [1, p. 286] it now follows that the distribution of

$$(3.3) \qquad (\xi,\eta) = \left(\frac{\tau_1 - 2n}{2\sqrt{n}},\ \frac{\tau_2 - 2n}{2\sqrt{n}}\right)$$

tends to the bivariate normal distribution given by the frequency function

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}}\, e^{-\frac{1}{2(1-\varrho^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2\varrho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)}$$

$$(3.4) \qquad = \frac{3}{\pi^2\sqrt{12-\pi^2}}\, e^{-6\frac{3x^2+(\pi^2-6)xy+3y^2}{\pi^2(12-\pi^2)}}$$

as $n \to \infty$. Some important fractiles of this asymptotic distribution are tabulated below.

(1)	$P(\xi > x,\ \eta \leq -x)$ } %	20.0	10.0	5.0	2.5	1.0	0.5	0.1	0.05
(2)	$P(\xi > x,\ \eta > x)$ } %								
(1)	x	0.46	0.89	1.20	1.50	1.84	2.07	2.55	2.73
(2)	x	–	0.10	0.30	0.46	0.65	0.76	1.01	1.11

Table 1. Some fractiles of the bivariate normal distribution with means (0, 0), standard deviations (1, 1), and coefficient of correlation $1-\pi^2/6$.

(The table has been made by Klaus Illum, M. Sc., with the aid of an electronic computer.)

As

$$\mathcal{M}\{a\xi + b\eta\} = 0, \quad \mathcal{V}\{a\xi + b\eta\} = a^2 + b^2 + 2\left(1 - \frac{\pi^2}{6}\right)ab\,,$$

we further have the asymptotic probability

$$(3.5) \qquad P(a\xi + b\eta \leqq x) = \Phi\left(\sqrt{\frac{3}{3(a^2+b^2)+(6-\pi^2)ab}}\,x\right),$$

where Φ is the normal distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\,dt\,.$$

4. Example. The exponential distribution. Consider the exponential distribution

$$F(x) = \begin{cases} 0 & \text{for} \quad u < -1 \\ 1-e^{-(u+1)} & \text{for} \quad u \geqq -1\,, \end{cases}$$

where u is the standardized variable $u = \dfrac{x-m}{\sigma}$.

From a given sample $x_1, x_2, \ldots, x_n$ it is desired to test the hypothesis H_0: $m=m_0$, $\sigma=\sigma_0$ against the hypothesis H_1: $m=m_1$, $\sigma=\sigma_1$.

First it may be remarked that if some $u_i=(x_i-m_0)/\sigma_0$ in the sample are less than -1, the hypothesis H_0 is false with probability 1. Assuming all u_i greater than -1, formula (3.2) gives

$$(4.1) \quad \tau_2 = -2\sum_{i=1}^{n} \ln[1-F(x_i)] = 2\sum_{i=1}^{n}(u_i+1) = 2n\left(\frac{\bar{x}-m_0}{\sigma_0}+1\right),$$

where $n\bar{x}=\sum_{i=1}^{n} x_i$. If H_0 is true, τ_2 is distributed as χ^2 with $2n$ degrees of freedom.

Following Cramér [1, p. 529], we will try to find *a uniformly most powerful test* corresponding to a given probability level ε of rejecting H_0 when it is true.

Writing the joint frequency function of the sample as $f(\boldsymbol{x};\boldsymbol{\alpha})$, where $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ and $\boldsymbol{\alpha}=(m,\sigma)$, we have

$$f(\boldsymbol{x};\boldsymbol{\alpha}) = \frac{1}{\sigma^n} e^{-\sum_{i=1}^{n}\left(\frac{x_i-m}{\sigma}-1\right)} = \frac{1}{\sigma^n} e^{n\left(\frac{m}{\sigma}-1\right)}\, e^{-\frac{n\bar{x}}{\sigma}}$$

for all $x_i > m-\sigma$, and $f(\boldsymbol{x};\boldsymbol{\alpha})=0$ elsewhere.

The searched critical set for $\boldsymbol{x}$ in the sample space R^n is determined by the condition

$$(4.2) \qquad f(\boldsymbol{x};\boldsymbol{\alpha}_1) \geqq c\cdot f(\boldsymbol{x};\boldsymbol{\alpha}_0)\,,$$

where c is some positive constant. Only if $m_1-\sigma_1 \leqq m_0-\sigma_0$ is a simple critical set obtained, namely the set that follows from

(4.3) $$e^{\left(\frac{1}{\sigma_0}-\frac{1}{\sigma_1}\right)n\bar{x}} \geqq \text{constant} > 0\ ,$$

in addition to the set

(4.4) $$\{\boldsymbol{x} \mid \text{at least one } x_i < m_0-\sigma_0\}\ .$$

For $\sigma_1 > \sigma_0$, (4.3) is fulfilled with the probability ε, see (4.1), when

$$\tau_2 = 2n\left(\frac{\bar{x}-m_0}{\sigma_0}+1\right) \geqq \chi^2_{2n,(1-\varepsilon)},$$

or

$$n\bar{x} \geqq n(m_0-\sigma_0)+\frac{\sigma_0}{2}\chi^2_{2n,(1-\varepsilon)}.$$

The critical set found is independent of $\boldsymbol{\alpha}_1=(m_1,\sigma_1)$ and is thus the uniformly most powerful test of H_0 relative to all admissible H_1. Obviously, the test is *biased* as it is required that $m_1-\sigma_1 \leqq m_0-\sigma_0$ and $\sigma_1>\sigma_0$. For $m_1-\sigma_1 \leqq m_0-\sigma_0$ and $\sigma_0>\sigma_1$ we get the test (4.4) and

$$2n\left(\frac{\bar{x}-m_0}{\sigma_0}+1\right) \leqq \chi^2_{2n,\varepsilon}\ .$$

The probability measure of the set (4.4) is zero as long as the hypothesis H_0 is true but not if H_1 is true. Due to this, it becomes complicated to work out the power of the test unless we further assume that $m_1-\sigma_1=m_0-\sigma_0$. This last situation is the usual one in most applications of the exponential distribution, for instance in traffic theory. We then normally have

$$m_1-\sigma_1 = m_0-\sigma_0 = 0\ .$$

It is then easy to obtain the power function for the test, as the random variable

$$2n\left(\frac{\bar{x}-m_1}{\sigma_1}+1\right) = \tau_2\frac{m_0}{m_1}$$

is distributed as χ^2 with $2n$ degrees of freedom under the hypothesis H_1. The power function becomes

(4.5) $$P(\tau_2 \geqq \chi^2_{2n,(1-\varepsilon)}) = P\left(\chi^2_{2n} \geqq \chi^2_{2n,(1-\varepsilon)}\frac{m_0}{m_1}\right)$$

for the test of H_0 against H_1. Requiring that

(4.6) $$P(\tau_2 \geqq \chi^2_{2n,(1-\varepsilon)}) = \begin{cases}\varepsilon & \text{when } H_0 \text{ is true}\\ 1-\varepsilon & \text{when } H_1 \text{ is true}\ ,\end{cases}$$

we get the formula

$$\frac{\sigma_0}{\sigma_1} = \frac{m_0}{m_1} = \frac{\chi^2_{2n,\varepsilon}}{\chi^2_{2n,(1-\varepsilon)}}$$

for a significance bound of m_0/m_1 on level ε for errors of the first and second kind, when a sample of n members is given, see fig. 1.

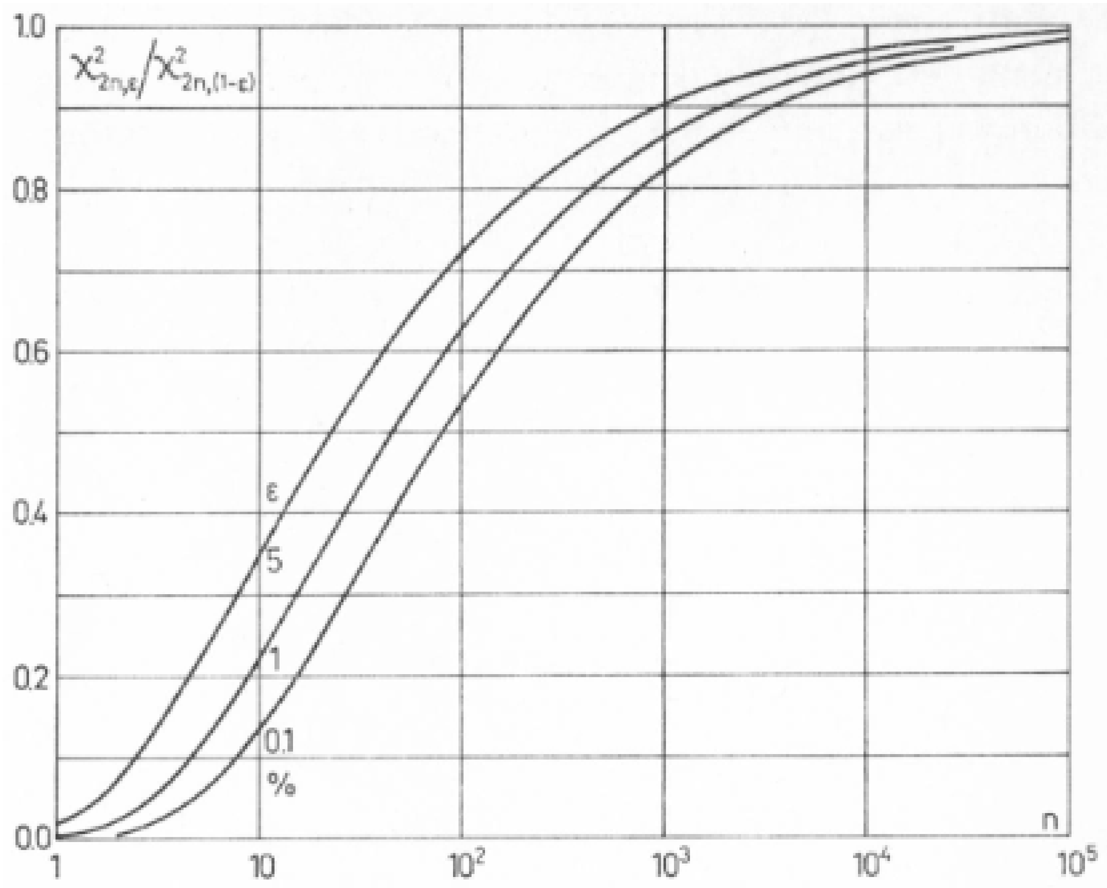

Fig. 1

5. Example. Pseudo-random numbers. The uniform distribution. If a stochastic experiment has to be simulated on a computer, a simple method to generate random numbers that are uniformly distributed on the interval [0, 1] must be at hand.

Random numbers may be generated from a stochastic physical process, but this matter is often difficult to handle. Normally, so-called pseudo-random numbers are used. They are, curiously enough, generated in a purely deterministic way, namely by a simple recursive mathematical formula, for instance by the additive or the multiplicative congruential method. For references, see [4], [5] and [6].

A method of the additive type is the Fibonacci sequence [4]. Starting with two random integers r_0 and r_1, we get a sequence of "random" integers by the formula

$$r_n \equiv r_{n-1} + r_{n-2} \pmod{M} . \tag{5.1}$$

After division by $M-1$, we get a sequence of "random" numbers that are uniformly distributed on the interval [0, 1].

Naturally, the applicability of such a sequence of numbers as representative for the outcome of a sequence of independent stochastic experiments must be tested with a great variety of tests. Such investigations have been published by various authors. They show that pseudo-random numbers generated by multiplicative and certain additive congruential methods follow the laws of probability in a satisfactory way for practical purposes, provided the period of the sequences is not surpassed.

The Fibonacci sequence (5.1) has the advantage of being very fast, but it has been asserted that it is unsatisfactory with respect to "randomness".

The object of this example is to test the Fibonacci sequence for uniformity, that is, we set up the hypothesis H_0 that the distribution functions is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leqq x < 1 \, . \\ 1 & \text{for } 1 \leqq x \, . \end{cases}$$

As alternative hypothesis H_1, it is simple to consider the beta distribution with the density function

$$\beta(x;p,q) = \begin{cases} \dfrac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1} & \text{for} \quad 0 \leqq x \leqq 1 \\ 0 & \text{for} \quad x < 0 \text{ and } x > 1 \, , \end{cases}$$

where $p>0$ and $q>0$ are parameters, and Γ is the gamma function. The two hypotheses coincide for $p=q=1$.

The condition (4.2) for the best test becomes

$$(p-1)\left(-2\sum_{i=1}^{n}\ln x_i\right)+(q-1)\left(-2\sum_{i=1}^{n}\ln(1-x_i)\right) \leqq -2\ln\left[c\left(\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}\right)^n\right]$$

$$= 2\sqrt{n}C+2n(p-1)+2n(q-1) \, , \tag{5.2}$$

or with (3.1), (3.2) and (3.3):

$$(p-1)\xi+(q-1)\eta \leqq C \, .$$

The probability of this event is determined by (3.5) if n is large. As this probability depends on p and q, a *uniformly* most powerful test of H_0 with respect to all H_1 does not exist.

The most powerful critical set on level ε is, for n large,

$$a\xi+b\eta \leqq \frac{u_\varepsilon}{\sqrt{3}}\sqrt{3(a^2+b^2)+(6-\pi^2)ab} \, , \tag{5.3}$$

where $a=p-1$, $b=q-1$, and u_ε is the ε-fractile of the standardized normal distribution.

The power function is only obtainable in a simple way if p or q are equal to 1. If, for example, $q=1$, the distribution function of the beta distribution is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^p & \text{for } 0 \leqq x < 1 \\ 1 & \text{for } 1 \leqq x \,, \end{cases}$$

and if H_1 is true,

$$p\tau_1 = -2\sum_{i=1}^{n} \ln x_i^p$$

is distributed like χ^2_{2n}. If $p>1$, (5.2) shows that $\tau_1 \leqq \chi^2_{2n,\varepsilon}$ is the best critical set, such that the power function becomes

$$P(\tau_1 \leqq \chi^2_{2n,\varepsilon}) = P(\chi^2_{2n} \leqq p\chi^2_{2n,\varepsilon}) \,.$$

If $p<1$, the power function becomes

$$P(\tau_1 \geqq \chi^2_{2n,(1-\varepsilon)}) = P(\chi^2_{2n} \geqq p\chi^2_{2n,(1-\varepsilon)}) \,,$$

which is the same as (4.5). With the requirement (4.6), we get the same curves as in fig. 1.

The Fibonacci sequence (5.1) has been tested on the Danish computer GIER, with $r_0=394852741$, $r_1=263822912$, $M=2^{29}-1=536870911$, and $n=10000$. As M is the largest (single word length) integer that can be stored in GIER, a small rounding off error will occur. This is, however, only an insignificant "random" contribution to r.

For every 100 experiments, (ξ,η) was calculated from (3.3) and plotted in the coordinate system shown in fig. 2. Some frequencies obtained from the 100 points are given in table 2, together with theoretical probabilities taken from the table 1. Significance lines corresponding to (5.3), with $a=b$ and $\varepsilon=5$, 1 and 0.1%, are drawn in fig. 2. No significant departure from the hypothesis H_0 is observed.

Probability %	1	5	10	13	20	36
1. quadrant	0	2	5	6	–	–
2. quadrant	2	3	8	–	22	35
3. quadrant	2	11	17	24	–	–
4. quadrant	1	3	10	–	21	35

Table 2. Observed distribution of 100 points (ξ, η) (see fig. 2).

From all 10000 experiments, (ξ,η) was calculated to be

$$(\xi,\eta) = (-0.1948,\ -0.4921)\,,$$

which is a very probable result under the hypothesis H_0.

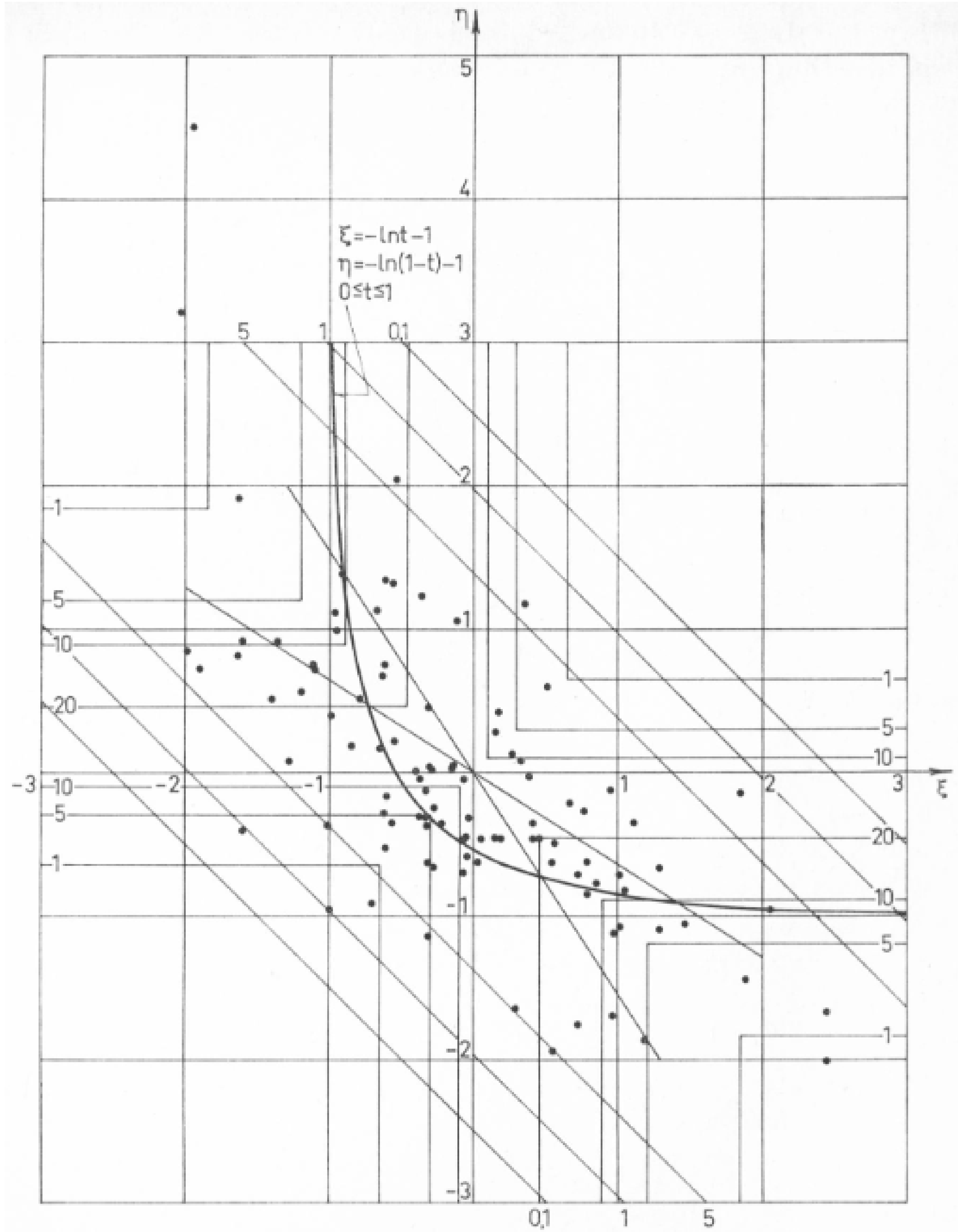

Fig. 2. 100 points each generated after (3.3) for $n = 100$. Thus every term in this sum may be interpreted as a point distributed at "random" on the curve $\xi = -\ln t - 1$, $\eta = -\ln(1-t) - 1$, where $0 \leqq t \leqq 1$. The parameter was drawn as $t = r$ by the reduced Fibonacci sequence. Domains with given probabilities (in %) corresponding to the asymptotic normal distribution (3.4) are indicated. Further, the lines of regression are drawn.

Further, estimates were calculated for the standard deviations of $\tau_1/2$ and $\tau_2/2$ and the coefficient of correlation between τ_1 and τ_2 by the usual formulae. The results were

$$\sigma\{\tau_1/2\} \approx 1.0041, \ \sigma\{\tau_2/2\} \approx 0.9916, \ \varrho\{\tau_1, \tau_2\} \approx -0.6398 ,$$

which are in good agreement with the theoretical quantities, see p. 160.

The distribution of the 10000 numbers in 10 equal intervals was

$$948,\ 1049,\ 1032,\ 953,\ 1004,\ 991,\ 1041,\ 1019,\ 982,\ 981\ .$$

The usual estimate of the variance of these uniformly distributed numbers r was (using the notation of standard errors)

$$\sigma^2\{r\} \approx 0{,}0827 \pm 7\cdot 10^{-4}\ .$$

The theoretical value is $\frac{1}{12} \approx 0.0833$. To get an impression of the dependence between numbers following each other in the sequence, the following coefficients of correlation were estimated:

$$\varrho\{r_n, r_{n+i}\} \approx \sum_{n=1}^{9996} (r_n - \tfrac{1}{2})(r_{n+i} - \tfrac{1}{2}) \Big/ \sum_{n=1}^{9996} (r_n - \tfrac{1}{2})^2\ ,$$

for $i = 1, 2, 3, 4$. The respective results were

$$-0.0033,\ -0.0108,\ -0.0006,\ -0.0048\ .$$

Added in the proof:

The author has recently become aware of the fact that the use of (2.2) and (2.3) as test statistics is classical. The method was first considered by R. A. Fisher (1932), K. Pearson (1933, 1934), J. Neyman (1937) and E. S. Pearson (1938), see for example Biometrica 30 (1938), p. 134.

REFERENCES

[1] H. Cramér: *Mathematical methods of statistics.* Princeton 1945.

[2] E. J. Gumbel: *On the reliability of the classical Chi-square test.* Annals of Math. Stat. 14 (1945), pp. 253–263.

[3] *Mathematical tables from the Handbook of chemistry and physics.* Eleventh edition. Chemical Rubber Publishing Co., 1959.

[4] O. Taussky — J. Todd: *Generation and testing of pseudo-random numbers.* Symposium on Monte Carlo Methods, pp. 15–28. John Wiley & Sons, New York 1956.

[5] Aimo Törn: *Tillämpning av Monte Carlo-metoden på beräkning av flerdimensionella integraler.* Nordisk Matematisk Tidskrift 11 (1963), pp. 97–102.

[6] K. T. Tocher: *The art of simulation.* The English Universities Press, London 1963.