CERN - Data Handling Division
DD/76/13
J.M.F. Chamayou
17 September 1976

[REDACTED]
CERN LIBRARIES, GENEVA

CM-P00059699

ALGORITHMS FOR THE KAC AND RENYI TESTS

(Submitted to Computer Physics Communications)

5
2
67

PROGRAM SUMMARY

Titles of programs: RENYIF, RENYIT, TESTKAC
Catalogue number:
Program obtainable from: CPC Program Library, Queen's University of Belfast,
N. Ireland.
Computer: CDC 7600; Installation CERN, Geneva.
Operating system: CDC Scope
Programming language: FORTRAN IV
High speed store required: 143 (RENYIF), 415 (RENYIT), 410 (TESTKAC)
No. of bits in a word: 60
Overlay structure: No
No. of magnetic tapes required: None
Other peripherals used: Card reader, line printer.
No. of cards in combined program test deck: 34 (RENYIF), 73 (RENYIT),
79 (TESTKAC)
Card punching code: ANSI
Keywords: Goodness of fit tests, asymptotic Renyi distribution, Renyi test,
Random size sample test, KAC test.
Nature of physical problem: Given a sample (with fixed size for the Renyi test
and random size for the Kac test) of independent
observations of a random variable having an unknown
continuous distribution function $G(x)$. The
hypothesis $G(x) = F(x)$ for a given $F(x)$ is tested
using the empirical distribution function.
Method of solution: The observations are ranked in increasing order to calculate
the empirical distribution function. The maximum of the
absolute difference between $F(x)$ and the empirical distribu-
tion function is computed and the test of the hypothesis is
performed using the asymptotic Renyi distribution.
Restriction on the complexity of the problem: The size of the sample must be
large enough to a valid use of the
asymptotic Renyi distribution.
Running time: Depends on the size of the sample and on the algorithm used to
rank the observations in increasing order.

LONG WRITE UP

1. INTRODUCTION

Let X be a random variable with the unknown continuous distribution function $G(x)$. We consider two different statistical tests for the simple hypothesis

$$H_0: G = F(x)$$

where F is a completely determined (i.e. independent of parameters to be estimated) continuous distribution function.

1.1 The Renyi test of goodness-of-fit

It is assumed that we are given n independent observations $X_1, X_2, \dots, X_n$ from X . Renyi [1] proposed the following test statistic for H_0 :

$$D_n = \sup_{0 < a \leq F(x) \leq b \leq 1} \frac{|F(x) - S_n(x)|}{F(x)} \quad (1)$$

where $a \neq b$ are arbitrary constants and

$$S_n(x) = \frac{1}{n} \sum_{i=1}^n h(x - X_i) \quad (2)$$

with

$$h(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 1 \end{cases} \quad (3)$$

is the empirical distribution function.

The distribution function of D_n for n large ($n \geq 50$) can be presented using the distribution function $L_1(z)$. $L_1(z)$ is the limiting distribution function of

$$D_n^* = \sqrt{n} D_n$$

for $n \rightarrow \infty$, which is given, if H_0 is true, by [1]:

$$L_1(z) = P(\lim_{n \rightarrow \infty} D_n^* \leq z) = L\left(\sqrt{\frac{a}{1-a}} z\right) \left(2\phi\left(\sqrt{\frac{b}{1-b}} z\right) - 1\right) + P_1(z) \quad (4)$$

where $L(z)$ is the asymptotical Renyi distribution function:

$$L(z) = \begin{cases} \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp \left\{ -((2k+1)\pi/2z)^2/2 \right\}, & \text{for real arguments } z > 0 \\ 0 & \text{for } z \leq 0 \end{cases} \quad (5)$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\xi^2/2} d\xi \quad (6a)$$

and

$$P_1(z) = \frac{4}{\pi} \left\{ \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp \left\{ -\frac{(2k+1)^2 \pi (1-a)}{8az^2} \right\} \cdot \int_0^{(2k+1)\pi/2} \exp \left\{ \frac{(1-b)\xi^2}{2bz^2} \right\} \sin \xi d\xi \right\} \cdot \frac{2 \exp \{-bz^2/2(1-b)\}}{\sqrt{2\pi} z \sqrt{\frac{b}{1-b}}} \quad (6b)$$

The test procedure finally runs as follows. We calculate $D_n^* = \sqrt{n} D_n$ for the given observations and we compute $L_1(D_n^*)$. If $L_1(D_n^*) < 1 - \alpha$, where α is the level of significance (e.g. $\alpha = 0.1, 0.05 \dots$) we accept H_0 otherwise we reject H_0 .

1.2 The Kac test of goodness-of-fit

In this case we assume that $N_\lambda, X_1, X_2, \dots$ are independent random variables where N_λ has a Poisson distribution with known mean value $\lambda > 0$:

$$P(N_\lambda = n) = e^{-\lambda} \frac{\lambda^{n-1}}{(n-1)!}, \quad n = 1, 2, \dots \quad (7)$$

and $X_1, X_2, \dots$ are observations from X for which the hypothesis H_0 has to be tested. Kac [2] proposed the following test statistic for H_0 :

$$D_{N_\lambda} = \sup_{0 \leq a \leq F(x) \leq b \leq 1} |F(x) - S_{N_\lambda}(x)| \quad (8)$$

where $a \neq b$ are arbitrary constants and

$$S_{N_\lambda}(x) = \frac{1}{\lambda} \sum_{i=1}^{N_\lambda} h(x - x_i) \quad (9)$$

with $h(x)$ given by (3), is a modified empirical distribution function ($S_{N_\lambda}(x)$ can naturally be greater than 1). If H_0 is true, the distribution function of D_{N_λ} for finite λ is given in a table by Allen and Beekman [3]. For n large ($n \geq 50$) one

can, however, use the distribution function $L_2(z)$ of the limiting value of

$$D_{N_\lambda}^* = \sqrt{\lambda} D_{N_\lambda}$$

for $\lambda \rightarrow \infty$, given by Csörgö [4]:

$$L_2(z) = P(\lim_{\lambda \rightarrow \infty} D_{N_\lambda}^* \leq z) = L\left(\frac{z}{\sqrt{b}}\right) \left[2\phi\left(\frac{z}{\sqrt{a}}\right) - 1 \right] + P_2(z) \quad (10)$$

where L and ϕ are given by (5) and (6), respectively, and

$$P_2(z) = \frac{4}{\pi} \left\{ \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp \left\{ -\frac{(2k+1)^2 \pi^2 b}{8z^2} \right\} \int_0^{(2k+1)\pi/2} \exp \{ a\xi^2/2z^2 \} \sin \xi \, d\xi \right\} \cdot \frac{2\sqrt{a} \exp \{-z^2/2a\}}{\sqrt{2\pi} z} \quad (11)$$

The following example may illustrate the use of the Kac test of goodness-of-fit:

AN EXAMPLE FROM PHYSICS

To illustrate the Kac test, we present the following example:

A radioactive source, whose known activity is of ξ curie, is used to perform, with the monoenergetic particles emitted, the spectrometry of the energy losses of these particles in a given medium. The spectrometry is performed during y seconds, (y being a short time compared to the time used to compute the activity). The number of particles detected during the time y will then be a random number distributed according to the Poisson distribution with a parameter λ . (This is the classical distribution of a radioactive decay). The mean λ will depend upon the activity ξ of the source and the duration y of the irradiation, (and other physical considerations related to the experiment such as: solid angle used to irradiate the medium, efficiency of the spectrometer and so on ...).

If the physicist wants now to test $F(x)$ the theoretical energy distribution against the experimental distribution of the pulses given by the spectrometer, he has a sample of energies: x_i , $i = 1, 2, \dots, n$, where n is the observed value of a Poisson variable N_λ with mean λ depending on the activity ξ and the duration y of the irradiation, which is the typical situation of the Kac test. The parameter λ must be large enough to a valid use of the asymptotic Renyi distribution.

Note: (In this example N_λ is clearly independent of the distribution of the observations x_i).

2. METHOD OF COMPUTATION

2.1 The asymptotical Renyi distribution function

To evaluate the series in (5), the range of the function $L(z)$ is divided into four parts:

for $z \leq 0.05$, $L(z) \equiv 0$ is taken ,

for $0.05 < z \leq 1$, the terms with $k \geq 5$ are neglected in (5) ,

for $z > 7.7$, $L(z) \equiv 1$ is taken ,

for $1 < z \leq 7.7$ the following formula, given by Takacs [5] is used:

$$L(z) = \sum_{k=-\infty}^{+\infty} (-1)^k \{ \phi((2k+1)z) - \phi((2k-1)z) \} \quad (12)$$

where $\phi(z)$ is given by (6). In (12) all terms with $k \leq -4$ and $k \geq 5$ are neglected.

2.2 The tests

For both tests we have to provide:

- the ranking of the observations X_i
- the value of $S_n(x)$ and S_{N_λ} , respectively, at the observations X_i
- the values $F(X_i)$
- the test statistics D_n^* and $D_{N_\lambda}^*$, respectively
- the distribution function values $L_1(D_n^*)$ and $L_2(D_{N_\lambda}^*)$, respectively.

The remainder term $P_1(z)$ in $L_1(z)$ is not computed exactly but use is made of the bound:

$$\frac{2 \exp \{ -bz^2/2(1-b) \}}{\sqrt{2\pi} z \sqrt{\frac{b}{1-b}}} L \left(z \sqrt{\frac{ab}{b-a}} \right) \geq P_1(z) \quad (13)$$

since $P_1(z)$ is very small for b sufficiently close to 1. In the same way the remainder term $P_2(z)$ in $L_2(z)$ is replaced by the bound

$$\frac{2\sqrt{a} \exp \{ -z^2/2a \}}{\sqrt{2\pi} z} L \left(\frac{z}{\sqrt{b-a}} \right) \geq P_2(z) \quad (14)$$

since $P_2(z)$ is very small for a sufficiently close to zero. Both bounds are computed separately in the programs such that the user can have an idea of the size of $P_1(z)$ and $P_2(z)$, respectively.

3. USAGE

3.1 The asymptotical Renyi distribution function

The function subprogram RENYIF(Z) computes the value of $L(z)$ for $Z = z$, where L is defined by (5). The function $\phi(z)$ is called by RENYIF under the name $FREQ(Z)$. This routine has to be supplied by the user. Such a routine is, for instance, provided by Cooper's algorithm AS2 [6].

The precision was of 12 exact decimal places for $L(z)$ when RENYIF was performed on a CDC 7600.

3.2 The tests

i) The Renyi test of goodness of fit

The subroutine RENYIT (FF, N, F, K, E, AA, BB, DN, Z, P, P1) calculates:

DN: Given by (1)

Z : $\text{SQRT}(N) * DN$

P : Given by (4) where $P_1(z)$ has been replaced by the bound (13)

P1: Given by (13)

The user has to provide:

N: The sample size

AA = a and BB = b, [see (1)]

An error message is printed when AA and BB are illegal [e.g. $AA \leq 0$, $BB \geq 1$, $AA \geq BB$, $F(X_N) \leq AA$]

F : An N dimensional array containing the N observations given by the user. This array is destroyed and returns as output the values of the distribution function F for each of the N ranked observations. (Note the double use of F, to avoid confusion use is made in the sequel of F^* when F represents the input i.e. the observations).

FF: User supplied function subroutine with two arguments defined in the following way by the user provided integer K:

K = 1: Corresponds to the case where the distribution function $F(x)$ is not known analytically but has to be calculated by numerical integration of the corresponding probability density function $f(x)$. In this case the user has to provide an external function FF where:

$$FF(X,Y) = \int_Y^X f(\xi) d\xi \quad (= F(X) - F(Y))$$

and

$$E \leq Y < X ,$$

where

E: left end point of the interval in which the observations are supposed to vary. This way we avoid to perform the whole integration of $f(\xi)$ for each observation.

K = 2: Corresponds to the case where the distribution function $F(x)$ is known analytically. In this case the user has to provide an external function FF, where

$$FF(X,E) = \int_{-\infty}^X f(\xi) d\xi \quad (= F(X))$$

and $E = 0$ for convenience since it is actually not used in this case.

Remark 1. Given the N-dimensional array F^* of observations, the subroutine FLPSOR(F^*, N) is used to rank these values in increasing order. This routine has to be supplied by the user. Such a routine is, for instance, provided by Scowen's algorithm 271 [7].

Remark 2. The values $F(I)$ of the distribution function at the points $F^*(I)$ are computed with the aid of the user supplied external function FF:

$$\begin{aligned} \text{if } K = 1 , \quad F(I) = F(F^*(I)) &= \sum_{J=1}^I FF(F^*(J) , F^*(J-1)) = \\ &= F(I-1) + FF(F^*(I) , F^*(I-1)) = \int_E^{F^*(I)} f(\xi) d\xi , \quad I = 2, 3, \dots, N \end{aligned}$$

where

$$F(1) = FF(F^*(1) , E)$$

and

$$F^*(0) = E$$

$$\text{if } K = 2 , \quad F(I) = F(F^*(I)) = FF(F^*(I) , E) = \int_{-\infty}^{F^*(I)} f(\xi) d\xi , \quad I = 1, 2, \dots, N$$

Remark 3. The function FREQ(Z) has to be provided by the user, with the same definition as in RENYIF.

ii) The Kac test of goodness of fit

The subroutine TESTKAC (FF, N, F, K, E, AA, BB, AL, M, DN, Z, P, P2) calculates:

DN: Given by (8)

Z : $\text{SQRT}(\text{AL}) * \text{DN}$

P : Given by (10) where $P_2(z)$ has been replaced by the bound (14)

P2: Given by (14)

The user has to provide:

N : The observed value of the Poisson random size of the sample

M : The dimension of the array F, ($M \geq N$)

AL = λ the mean of the Poisson random size of the sample

AA = a and BB = b, [see (8)]

An error message is printed when AA and BB are illegal [e.g. $\text{AA} \leq 0$, $\text{BB} \geq 1$, $\text{AA} \geq \text{BB}$, $F(X_N) \leq \text{AA}$]

F : An M dimensional array containing the N observations, given by the user.
This array is destroyed and returns as output the values of the distribution function F for each of the N ranked observations.

FF, K and E are defined in the same way as in RENYIT

Remarks 1, 2, 3 hold as in RENYIT.

Remark 4. The minimal value of AL, the mean of the sample size, is left to the user appreciation but it is recommended, according to the results in [3], to use the asymptotical Renyi distribution function only for $\text{AL} \geq 50$.

Remark 5. I hope that the three algorithms presented here will develop the use of the Kac and Renyi tests in the applications in concurrence with the already available algorithms for the Kolmogorov-Smirnov and Cramer-Von Mises tests, an example of application can be found in [8].

TEST RUN

RENYIF is tested using 2000 gaussian random numbers from [9] for $a = 0.05$ and $b = 0.95$.

TESTKAC is tested using N gaussian random numbers for $\lambda = 50$, N is given by the Poisson random number generator subroutine POISSN (AL, L, IER) from the CERN computer Library (multiplicative method).

Acknowledgement

The author is deeply indebted to Dr. B. Schorr for making helpful comments.

Table of the Asymptotical RENYI Distribution

Z	L(Z)
0.2	0.5 10-13
0.3	0.14 10-5
0.4	0.00057
0.5	0.00916
0.6	0.04136
0.7	0.10267
0.8	0.18524
0.9	0.27761
1.0	0.37078
1.1	0.45927
1.2	0.54036
1.3	0.61299
1.4	0.67703
1.5	0.73278
1.6	0.78081
1.7	0.82174
1.8	0.85628
1.9	0.88513
2.0	0.90900
2.1	0.92854
2.2	0.94439
2.3	0.95710
2.4	0.96721
2.5	0.97516
2.6	0.98136
2.7	0.98613
2.8	0.98978
2.9	0.99254
3.0	0.99460
3.5	0.99907
4.0	0.99987

REFERENCES

- [1] A. Renyi, On the theory of order statistics, Acta Math. Acad. Sci. Hungar. 4, 191-231 (1953).
- [2] M. Kac, On deviations between theoretical and empirical distributions, Proc. Nat. Acad. Sci., U.S.A., 35, 252-257 (1949).
- [3] J.F. Allen and J.A. Beekman, Distribution of a M. Kac statistics, Ann. Math. Statist. 38, 1919-1923 (1967).
- [4] M. Csörgö, Distribution results for distance functions based on the modified empirical distribution function of M. Kac, Ann. Inst. Math. Stat. 24, 101-110 (1972).
- [5] L. Takacs, On random walk problems, Pub. Math. Inst. Hung. Acad. Sci., 2, 81-90 (1957).
- [6] B.E. Cooper, The normal integral, Appl. Statist. 17, 186-188 (1968).
- [7] R.S. Scowen, "Quicksort", Algorithm 271, Com. Ass. Comp. Mach. 8, 669-670 (1965).
- [8] J.M.F. Chamayou, On a direct algorithm for the generation of log-normal pseudo-random numbers, Computing 16, 69-76 (1976).
- [9] A million random digits; Glencoe, Ill.: The Free Press, The Rand Corporation 1955.