

TEST OF INDEPENDENCE FOR DISCRETE DISTRIBUTIONS BASED ON THE EMPIRICAL GENERATING FUNCTION

consistency/Empirical generating function/measure of dependence/von Mises's statistics/U-statistics

JAMAL-DINE CHERGUI

B. P.: 706. Tetouan, Morocco.

Presentado por Francisco J. Girón González-Torre, 10 de Abril de 1996

ABSTRACT

In 1948 Hoeffding, W ([7]) proposed the functional

$$\Delta = \int \left(F_{(X,Y)}(x,y) - F_1(x)F_2(y) \right)^2 dF_{(X,Y)}(x,y)$$

based on the distance between the joint and marginal distribution functions for measuring dependence. This functional isn't fully satisfactory. It's an appropriate measure only when F is absolutely continuous. In this article we suggest the functional based on the generating function for measuring dependence and testing independence for discrete distributions. The corresponding empirical functional is essentially the statistics to be considered for testing independence.

RESUMEN

En 1948 Hoeffding, W ([7]) propuso el funcional

$$\Delta = \int \left(F_{(X,Y)}(x,y) - F_1(x)F_2(y) \right)^2 dF_{(X,Y)}(x,y)$$

basado sobre la distancia entre la función de distribución conjunta y el producto de las funciones de distribución marginales para medir la dependencia. Es una buena medida cuando F es absolutamente continua. En este artículo se propone un funcional basado sobre la función generatriz para medir la dependencia y testar la independencia de las distribuciones discretas.

1. INTRODUCTION

Measuring dependence and testing independence are one of the most important aspects of many statistical investigation.

This problem has been receiving considerable attention. In the independence problem we want to test if two

(or in general n) random variables X and Y with marginal distributions F_1, F_2 and bivariate distribution $F_{(X,Y)}$ are independent. This hypothesis of independence can be tested in a nonparametric framework.

"The idea of using various simple functionals of the sample d.f of vector chance variables in order to test the independence of components, is a natural one" ([1]). Many functionals have been proposed and studied in the statistical literature.

The following measure which is based on the distance between two distribution functions, is proposed by Hoeffding, W ([7])

$$\Delta = \int \left(F_{(X,Y)}(x,y) - F_1(x)F_2(y) \right)^2 dF_{(X,Y)}(x,y)$$

This function is not fully satisfactory as measure of dependence, since the examples of the discrete distributions may be found where $\Delta = 0$ in the presence of dependence. It's an appropriate measure when $F_{(X,Y)}$ is absolutely continuous.

Example 1. (Kumar Joag Dev ([8], p. 84))

If $P(X = 0, Y = 1) = P(X = 1, Y = 0) = 1/2$. It's easy to see that $\Delta = 0$. However, X and Y are dependent.

The main objective of this article is to measure dependence and test independence for the discrete distributions. Our attention is aimed at the use of the generating function. The discussion is limited only on the bivariate distributions, as its extension to multidimensional distributions is straightforward.

Let (X, Y) be a random variable defined on the probability space (Ω, \mathcal{A}, P) taking values in the measurable space $(\mathbb{N}^2, \mathcal{P}(\mathbb{N}^2))$.

Let G, G_1 and G_2 denote the generating functions of $(X, Y), X$ and Y respectively. We suggest the squared dif-

ference functional for measuring dependence and testing independence

$$I = \int_T (G(s, t) - G_1(s)G_2(t))^2 ds dt; T = [0, 1]^2$$

Let $(X_1, Y_1); \dots; (X_n, Y_n)$ be independent identically distributed copies of (X, Y) .

Let $G_n, G_{n,1}, G_{n,2}$ be the empirical generating function associated with the sample $\{(X_i, Y_i); 1 \leq i \leq n\}$, $\{X_i; 1 \leq i \leq n\}$ and $\{Y_i; 1 \leq i \leq n\}$ respectively, that is to say:

$$G_n(\cdot, (s, t)) = \frac{1}{n} \sum_{i=1}^n s^{X_i(\cdot)} t^{Y_i(\cdot)}$$

$$G_{n,1}(\cdot, s) = G_n(\cdot, (s, 1)), G_{n,2}(\cdot, t) = G_n(\cdot, (1, t))$$

G_n is a sufficient, strongly consistent unbiased estimator of the generating function G in such a way that $E_n = \sqrt{n}(G_n - G)$ converges in cylindrical law to $N_{C,M}(O, Q)$ ([4]). An intuitively appealing estimate of I is $I_n = \int_T T_n^2(s, t) ds dt$ where $T_n(s, t) = G_n(s, t) - G_{n,1}(s)G_{n,2}(t)$.

2. CONSISTENCY

Here we want to establish that I_n is a strongly estimator of I .

Proposition 1. We have:

a/ $EI_n \longrightarrow I$ as $n \rightarrow \infty$

b/ I_n converges almost surely as $n \rightarrow \infty$ to I

Proof:

$$T_n^2(s, t) = (n-1)^2/n^4 \left(\sum_{i=1}^n s^{X_i} t^{Y_i} \right)^2 - 2(n-1)/n^4 \left(\sum_{i=1}^n s^{X_i} t^{Y_i} \right) \left(\sum_{i \neq j}^n s^{X_i} t^{Y_i} \right) + \left(1/n^4 \right) \left(\sum_{i \neq j}^n s^{X_i} t^{Y_i} \right)^2$$

simple but somewhat tedious calculations which can be obtained by writing to the author, yield

$$\begin{aligned} ET_n^2(s, t) &= \left[a_n G(s^2, t^2) + b_n G^2(s, t) + c_n G(s, t)G_1(s)G_2(t) + \right. \\ &\quad \left. d_n \left(G(s^2, t)G_2(t) + G(s, t^2)G_1(s) \right) + e_n G_1^2(s)G_2^2(t) \right. \\ &\quad \left. + f_n \left(G_1(s^2)G_2^2(t) + G_1^2(s)G_2(t^2) \right) \right] / n^4 \end{aligned}$$

where $a_n = n^2(n-1)$; $b_n = n(n-1)^3$; $c_n = -2n(n-1)^2(n-2)$

$$d_n = -2n(n-1)^2; e_n = n(n-1)(n^2 - 3n + 3);$$

$$f_n = n(n-1)(n-2)$$

therefore, it follows that

$$EI_n = \int_T E(T_n^2(s, t)) ds dt \text{ tends to } I \text{ as } n \rightarrow \infty$$

We notice that

$$\begin{aligned} T_n(s, t) - T(s, t) &= G_n(s, t) - G_{n,2}(t)(G_{n,1}(s) - G_1(s)) \\ &\quad - G_1(s)(G_{n,2}(t) - G_2(t)) \end{aligned}$$

We obtain then

$$\|T_n - T\| \leq \|G_n - G\| + \|G_{n,1} - G_1\| + \|G_{n,2} - G_2\| \cdot \left(\left\| \cdot \right\| = \sup_T |\cdot| \right)$$

Since the right hand side tends to 0 a.s as $n \rightarrow \infty$ ([4]). We have

$$\|T_n - T\| \longrightarrow 0 \text{ a.s as } n \rightarrow \infty$$

$$\text{Let us write } I_n = I + \int_T (T_n^2(s, t) - T^2(s, t)) ds dt = I + A$$

Since $|A| \leq 4\|T_n - T\|$, it's clear that I_n converges almost surely to I as $n \rightarrow \infty$

3. ASYMPTOTIC DISTRIBUTION

Here we give the asymptotic distribution of I_n under the hypothesis H_0 : "X and Y are independent". The following result is needed.

Lemma 1. ([11] p. 4): Let $a(z_1, \dots, z_m)$; $z_i = (x_i, y_i)$, $1 \leq i \leq m$ be a bounded function from IR^{2m} into IR . Then

$$\int \dots \int a(z_1, \dots, z_m) dD_n(z_1) \dots dD_n(z_m) = 0_p(1) \quad (*)$$

where $D_n = \sqrt{n}(F_n - F)$ is the empirical process.

Proof. We use lemma B ([9]. p. 223). We have

$$E \left(\int \dots \int a(z_1, \dots, z_m) dD_n(z_1) \dots dD_n(z_m) \right)^2 = 0 \quad (1)$$

Using chebyshev's inequality, we obtain (*).

The following Proposition establishes our intuition that, under the null hypothesis of independence, I_n is asymptotically equal to a Cramer-von Mises statistics.

Proposition 2. Under the hypothesis H_0 we have:

$$I_n = \frac{1}{n^2} \sum_{i,j=1}^n h((X_i, Y_i); (X_j, Y_j)) + O_p(n^{-3/2})$$

$$\text{where } h((x_1, y_1); (x_2, y_2)) = \int_T q((s, t); (x_1, y_1)) q((s, t); (x_2, y_2)) ds dt$$

$$\text{with } q((s, t); (x, y)) = (s^x - G_1(s))(t^y - G_2(t)).$$

Proof. Write

$$T_n(s, t) = G_n(s, t) - G_1(s)G_2(t) - G_1(s)(G_{n,2}(t) - G_2(t)) - G_2(t)(G_{n,1}(s) - G_1(s)) - (G_{n,1}(s) - G_1(s))(G_{n,2}(t) - G_2(t)).$$

$$\text{Let's put } u = (s, t); \quad u = (x, y) \quad \text{and} \quad H_i(u, z) = (u^i)^z - G_i(u^i), \quad i = 1, 2$$

where $(\cdot)^i$ is the i -th component of (\cdot) .

Then it follows that $\int H_i(u, z) dF(z) = 0$ yields to

$$T_n(s, t) = \int H_1(u, z) H_2(u, z) dF_n(z) - \iint H_1(u, z_1) H_2(u, z_2) dF_n(z_1) dF_n(z_2) \\ = n^{-1} \int H_1(u, z) H_2(u, z) dD_n(z) - n^{-1} \iint H_1(u, z_1) H_2(u, z_2) dD_n(z_1) dD_n(z_2)$$

Using the lemma 1 we obtain

$$T_n^2(s, t) = n^{-1} \left(\iint \prod_{i,j=1,2} H_i(u, z_j) dD_n(z_1) dD_n(z_2) \right) + O_p(n^{-3/2})$$

$$\text{We define } q((s, t); (x, y)) = (s^x - G_1(s))(t^y - G_2(t))$$

$$h(z_1, z_2) = \int_T q((s, t); z_1) q((s, t); z_2) ds dt$$

Thus

$$I_n = \int_T T_n^2(s, t) ds dt = n^{-1} \iint \left(\int_T \prod_{i,j=1,2} H_i((s, t), z_j) ds dt \right) \\ dD_n(z_1) dD_n(z_2) + O_p(n^{-3/2})$$

$$I_n = \frac{1}{n^2} \sum_{i,j=1}^n h((X_i, Y_i); (X_j, Y_j)) + O_p(n^{-3/2})$$

which can be written in the form:

$$nI_n = nV_n + O_p(n^{-1/2})$$

nV_n is the von Mises' statistics which is associated with the Kernel h .

The Kernel h induces the integral operator by

$$Af(i, j) = \sum_{k,l \geq 0} h((i, j); (k, l)) f(k, l) d_{k,1} d_{l,2}$$

where

$$d_{k,i} = \frac{1}{k!} \left(\partial^k G_i(s) / \partial s^k \right)_{s=0}, \quad i = 1, 2$$

The associated eigenvalues characterize the asymptotic distribution of nI_n and the corresponding eigenfunctions are orthonormal. We note that by rearranging terms, we get

$$h(z_1, z_2) = h_1(z_1^1, z_2^1) h_2(z_1^2, z_2^2)$$

with

$$h_1(x, y) = \int_0^1 (s^x - G_1(s))(s^y - G_1(s)) ds$$

$$h_2(x, y) = \int_0^1 (t^x - G_2(t))(t^y - G_2(t)) dt$$

We define the integral operators

$$B_i f(l) = \sum_{k \geq 0} h_i(l, k) f(k) d_{k,i} \quad i = 1, 2$$

Now let (λ_K^i, Φ_K^i) $k = 1, 2, \dots$; be an eigenpair of B_i $i = 1, 2$ and let us put

$$\lambda_{jK} = \lambda_j^1 \lambda_K^2, \quad \Phi_{jK}(s, t) = \Phi_j^1(s) \Phi_K^2(t)$$

It then follows that $(\lambda_{jK}, \Phi_{jK})$ $j, K \geq 1$ is an eigenpair of A .

Thus, to solve the integral equation to obtain λ_{jK} we need only to solve the integral equations for λ_j^1 and λ_K^2 .

We have ([9]. p. 196)

$$E(\Phi_{ij}(X, Y) \Phi_{kl}(X, Y)) = \delta_{(i,j)(k,l)}, \quad E(\Phi_{ij}(X, Y)) = 0 \quad \forall i, j$$

$$E(h((X, Y); (X, Y))) = \sum_{i,j} \lambda_{ij} < \infty$$

Remark. 1: Let's put $\sigma_K(f) = f(k) d_{k,1}$, $f_k(j) = h_1(j, k)$ and $\sigma_j = \sigma_j(f_k)$. If X has a finite spectrum $\{1, \dots, m\}$, then the eigenvalues of B_1 are the eigenvalues of the matrix $\sum_{1 \leq i, j \leq m} (\sigma_{ij})$. (We have the same result if the spectrum is $\{x_1, \dots, x_m\}$).

Proposition 3. Under the hypothesis H_0 , we have

$$nI_n \xrightarrow{d} \sum_{i,j} \lambda_{ij} N_{ij}^2$$

where $(N_{ij}; i, j \geq 1)$ are iid $N(0,1)$.

Proof. We have

$$nI_n = nV_n + O_p\left(n^{-\frac{1}{2}}\right)$$

$$nV_n = \frac{1}{n} \sum_{i=1}^n h((X_i, Y_i); (X_i, Y_i)) + \frac{2}{n} \binom{n}{2} U_n$$

Where U_n is the U -statistics that is associated with the Kernel h .

We have

a/ By the strong law of large numbers

$$\frac{1}{n} \sum_{i=1}^n h((X_i, Y_i); (X_i, Y_i)) \xrightarrow{n \rightarrow \infty} E(h(X, Y); (X, Y)) = \sum_{i,j} \lambda_{ij} < \infty \text{ a.s.}$$

$$\text{b/ By th. 1 ([2], p. 4), } n U_n \xrightarrow{d} \sum_{i,j \geq 1} \lambda_{ij} (N_{ij}^2 - 1)$$

These results together prove the previous proposition.

4. TEST OF INDEPENDENCE

The problem under study is that for testing H_0 : “ X and Y are independent” against the alternative H_1 . Choose a possible prescribed level of significance α ($0 < \alpha < 1$). We consider the following test: reject H_0 if $nI_n > u_\alpha$ where u_α is chosen so that an approximate level of significance is achieved. u_α is the upper α -point in the limit null distribution of nI_n .

Example 2. Let $X \rightarrow B(p)$ and $Y \rightarrow B(u)$; $0 < p, u < 1$

[$B(\cdot)$ is the Bernoulli distribution]. We want to test the null hypothesis H_0 that the two variables X and Y are independent. We have $G_1(s) = ps + q$, $G_2(t) = ut + v$; $p + q = u + v = 1$.

In the light of proposition 3 and remark 1, it's easy to check that

$$nI_n \xrightarrow{d} \sigma^2(p, u) \chi_1^2 \text{ where } \sigma^2(p, u) = (pquv)/9$$

For α : $0 < \alpha < 1$ we reject H_0 if $nI_n > u_\alpha$ where $P(\sigma^2(p, u) \chi_1^2 > u_\alpha) = \alpha$

REFERENCES

1. Blum, J.R., Kiefer, J. & Rosenblatt, M. (1961), Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **32**, 485-98.
2. Carlstein, E. (1986), Asymptotic distribution theory for degenerate U -statistics for stationary sequences. *Mimeo Series 1704*. Dept. of statistics. Chapel Hill, North Carolina. U.S.A.
3. Chan, N.H., Tran, L.T. (1992), Nonparametric tests for serial dependence. *Jour. of time series analysis*. Vol. 13. N.º 1, 19-28.
4. Chergui, J.D. (1994), Estimation and tests of the discrete probability Law based on the empirical generating function (two dimensional case). *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales*. (Esp). Tomo LXXXVIII, cuaderno segundo-tercero. Madrid.
5. Cotterill, D.S., Csörgö, M. (1985), On the limiting distribution of and critical values for the Hoeffding, Blum, Kiefer, Rosenblatt independence criterion. *Statistics & Decisions* **3**, 1-48.
6. Gregory, G.G. (1977), Large sample theory for U -statistics and tests of fit. *The Annals of statistics*. Vol. 5, N.º 1, 110-123.
7. Hoeffding, W. (1948), A non-parametric test of independence. *Ann. Math. Statist.* **19**, 293-325.
8. Joag-Dev, K., Measure of dependence. In P.R. Krishnaiah and P.K. Sen, eds., *Handbook of statistics*, Vol. 4. Elsevier Science Publishers (1984) 79-88.
9. Serfling, R. (1980), *Approximation theorems of Mathematical statistics*. New York. Wiley.
10. Shorack, G. R. & Wellner, J. A. (1986), *Empirical processes with Applications to statistics*. New York. Wiley.
11. Skaug, H. J. (1993), The limit distribution of the Hoeffding statistic for test of serial independence. Report n.º 23. Dept. of Mathematics. Univ. of Bergen. Norway.