Testing functional connection between two random variables

Gyula O.H. Katona
Rényi Institute, Budapest, Hungary
ohkatona@renyi.hu

Dedicated to Professor
Yuri Vasilevich Prokhorov
for his 80th birthday

1 Introduction

Let ξ and η be two, not necessarily independent random variables. The goal of the present paper is to study the situation when one needs to decide if η is a (deterministic) function of ξ or not, by using many independent tests.

The probability of the event that $\xi = k$ and $\eta = \ell$ is $p_{k,\ell}$, the probability of ξ being k is $p_k = \sum_\ell p_{k\ell}$. Suppose that we have m tests. Let ξ_i (η_i) $(1 \le i \le m)$ be totally independent copies of ξ (η). We will study the probability $\Pr(\xi \to \eta, m)$ of the event that m experiments (mis)indicate that ξ (deterministically) determines η, that is, there are no i and j $(1 \le i, j \le m)$ such that $\xi_i = \xi_j, \eta_i \ne \eta_j$.

Of course, if η is really a function of ξ then $\Pr(\xi \to \eta, m) = 1$ for every m, otherwise it is an decreasing function of m. The most practical case is when the probabilities $p_{k,\ell}$ are constant. Then the probability $\Pr(\xi \to \eta, m)$ tends to 0 when $m \to \infty$. One could ask many questions in this case, for instance to study the rate of convergence as a function of the $p_{k,\ell}$'s, but we will be investigating another case, namely the one when the probabilities are very small.

In the rest of the paper a series of probability distributions will be considered, that is $p_{k,\ell}(n),\ p_k(n)$ where n tends to infinity. The number of possible

values of ξ and η are finite, but this is also increasing with n. One can easily see that the smaller probabilities require a larger m to give a counter-example for the functional connection. Therefore m is also supposed to depend on n. For the sake of convenience we will not denote this dependence.

Heuristic form of Theorem 1. *If the probabilities uniformly decrease and m is increasing faster than*

$$\frac{1}{\sqrt{\sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2}}$$

then a counter-example shows, with large probability, that η is not a function of ξ. On the other hand, if m is increasing slower than the quantity above then the probability of a counter-example is nearly 0.

It is more convenient to use a logarithmic form in the precise formulation, this is why we introduce the following quantity:

$$H_2(\xi \to \eta) = -\log_2 \left(\sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2 \right). \tag{1}$$

Since the probabilities depend on n, the quantity $H_2(\xi \to \eta)$ will also do so (without denoting this dependence).

2 The statement

Let $p(\xi, \eta, I)$ denote the probability of the event that the pair $(\xi_1, \eta_1), (\xi_2, \eta_2)$ gives a counter-example, that is, $\Pr(\xi_1 = \xi_2, \eta_1 \neq \eta_2)$.

Similarly $p(\xi, \eta, V)$ denotes the probability of the event that the triple $(\xi_1, \eta_1), (\xi_2, \eta_2), (\xi_3, \eta_3)$ gives two counter-examples in the following way: $\xi_1 = \xi_2 = \xi_3, \eta_1 \neq \eta_2 \neq \eta_3$.

Finally $p(\xi, \eta, N)$ is the probability of the event that the quadruple $(\xi_1, \eta_1), (\xi_2, \eta_2), (\xi_3, \eta_3), (\xi_4, \eta_4)$ gives three counter-examples forming a path: $\xi_1 = \xi_2 = \xi_3 = \xi_4, \eta_1 \neq \eta_2 \neq \eta_3 \neq \eta_4$.

Theorem 1 *Suppose that*

$$\frac{p(\xi, \eta, V)^2}{p(\xi, \eta, I)^3} \to 0$$

and

$$\frac{p(\xi,\eta,N)}{p(\xi,\eta,I)^2}\to 0$$

hold. Then

$$\Pr(\xi\to\eta,m)\to\begin{cases}0 & \text{if } 2\log_2 m - H_2(\xi\to\eta)\to+\infty,\\ e^{-2^{a-1}} & \text{if } 2\log_2 m - H_2(\xi\to\eta)\to a,\\ 1 & \text{if } 2\log_2 m - H_2(\xi\to\eta)\to-\infty.\end{cases}$$

$p(\xi,\eta,I), p(\xi,\eta,V)$ and $p(\xi,\eta,N)$ will be expressed by the probabilities in the next section.

Motivations, consequences, related literature, and analysis of the conditions are postponed to the last section.

3 The proofs

Lemma 1 $p(\xi,\eta,I)=\sum_k p_k^2-\sum_{k,\ell}p_{k,\ell}^2$.

Proof. The left hand side is equal to $\Pr(\xi_u=\xi_v,\eta_u\neq\eta_v)$ by definition, what is equal to

$$\sum_k\sum_{\ell\neq\ell'}p_{k,\ell}p_{k,\ell'}=\sum_k\sum_{\ell,\ell'}p_{k,\ell}p_{k,\ell'}-\sum_k\sum_\ell p_{k,\ell}^2=$$

$$\sum_k\left(\sum_\ell p_{k,\ell}\right)^2-\sum_{k,\ell}p_{k,\ell}^2=\sum_k p_k^2-\sum_{k,\ell}p_{k,\ell}^2$$

□

Observe that $H_2(\xi\to\eta)=-\log_2 p(\xi,\eta,I)$.

Lemma 2

$$p(\xi,\eta,V)=\sum_k p_k^3-2\sum_{k,\ell}p_k p_{k,\ell}^2+\sum_{k,\ell}p_{k,\ell}^3$$

Proof. Use the simple sieve for the "space" $\xi_1=\xi_2=\xi_3$.

$$p(\xi,\eta,\mathrm{V})=\Pr(\xi_1=\xi_2=\xi_3,\eta_1\neq\eta_2\neq\eta_3)=$$

$$\Pr(\xi_1=\xi_2=\xi_3)-\Pr(\xi_1=\xi_2=\xi_3,\eta_1=\eta_2)-\Pr(\xi_1=\xi_2=\xi_3,\eta_2=\eta_3)+$$

$$\Pr(\xi_1=\xi_2=\xi_3,\eta_1=\eta_2=\eta_3)=$$
$$\Pr(\xi_1=\xi_2=\xi_3)-2\Pr(\xi_1=\xi_2=\xi_3,\eta_1=\eta_2)+$$
$$\Pr(\xi_1=\xi_2=\xi_3,\eta_1=\eta_2=\eta_3)=$$
$$\sum_k p_k^3-2\sum_k \Pr(\xi_3=k)\Pr(\xi_1=\xi_2=k,\eta_1=\eta_2)+\sum_{k,\ell} p_{k,\ell}^3$$

□

Lemma 3

$$p(\xi,\eta,N)=\sum_k p_k^4-3\sum_{k,\ell} p_k^2p_{k,\ell}^2+2\sum_{k,\ell} p_k p_{k,\ell}^3+\sum_k\left(\sum_\ell p_{k,\ell}^2\right)^2-\sum_{k,\ell} p_{k,\ell}^4.$$

Its proof is analogous to that of Lemma 2. □

Let $C_{k,\ell}$ $(1\le k,\ell\le m)$ be a partition of the set $\{1,2,\dots,m\}$, where some classes can be empty. The partition is denoted by $\mathcal{C}$. The vertex set of the graph $G(\mathcal{C})$ is $\{1,2,\dots,m\}$, two vertices x and y are joined by an edge if $x\in C_{k,\ell},\ y\in C_{k,\ell'}$ holds for some $\ell\neq\ell'$. Define $C_k=\cup_\ell C_{k,\ell}$, and let $|C_k|=c_k$. The subgraph of $G(\mathcal{C})$ induced by C_k is called a component even in the case when it is an empty graph (that is $C_{k,\ell}$ are empty for all ℓ with one exception). Suppose that $|C_{k,1}|\ge|C_{k,2}|\ge\ldots\ge|C_{k,m}|$.

A subgraph consisting of vertex-disjoint edges of a graph is a *matching* in G. The vertex-disjoint union of a matching and one path consisting of two edges is called a *V-matching*. Finally, the vertex-disjoint union of a matching and one path consisting of three edges is an *N-matching*.

Lemma 4 *Let $G(\mathcal{C})$ be the graph defined above. Then*

$$\sum_{\substack{\text{matching of}\\ j\ \text{edges}}}(-1)^j+2\sum_{\text{V-matching}}1+\sum_{\text{N-matching}}1\ge 0 \qquad (2)$$

where the matchings, V-matchings and N-matchings are subgraphs of $G(\mathcal{C})$.

Proof. First suppose that $c_k > 2$ holds for at least one k. Let $\mathcal{M}_+$ ($\mathcal{M}_-$) denote the family of all matchings of $G(\mathcal{C})$ consisting of even (odd) number of edges. Furthermore, $\mathcal{V}$ and $\mathcal{N}$ denote the families of all V-matchings and N-matchings, respectively. We will give a mapping f from $\mathcal{M}_-$ to $\mathcal{M}_+ \cup \mathcal{V} \cup \mathcal{N}$.

Suppose that $M \in \mathcal{M}_-$ has two edges in one of the components. It is easy to see that $G(\mathcal{C})$ contains an edge joining endpoints of these edges. Add this edge to M. The so obtained set $f(M)$ of edges is in $\mathcal{N}$. If M contains at most one edge in every component and a C_k with $c_k > 2$ contains an edge e then add another edge to this component, having a common endpoint with e. The so obtained set $f(M)$ of edges is in $\mathcal{V}$. Finally, suppose that every component contains at most one edge of M, but the components C_k with $c_k > 2$ none. Then add an edge of such a C_k with the smallest index. The so obtained $f(M)$ contains an even number of edges, therefore $f(M) \in \mathcal{M}_+$ holds.

The mapping f is not an injection, but "almost". If $M' \in \mathcal{N}$ then the middle edge of the path is uniquely determined, $|f^-(M')| = 1$. On the other hand, if $M' \in \mathcal{V}$ then M' could be obtained in two different ways, therefore $|f^-(M')| \leq 2$. Finally, if $M' \in \mathcal{M}_+$ then the new edge can be only in the component having the smallest index k with $c_k > 2$. Then, $|f^-(M')| = 1$ holds, again. The mapping f indirectly associates a $+1$ term with every -1 on the left hand side of (1), since the terms associated with the V-matchings are doubled. This proves the inequality in this case.

The only remaining case is when $c_k = 2$ ($1 \leq k \leq r$). If $|C_k| = |C_{k,\ell}|$ holds for some ℓ then this component contains no edge, it plays no role in (1). Therefore one can suppose that $|C_{k,1}| = |C_{k,2}| = 1$ holds for every k. Let the number of components with ta least one edge be r. Then $G(\mathcal{C})$ has r vertex-disjoint edges. It contains neither a V-matching nor an N-matching. The number of matchings M of j edges in $G(\mathcal{C})$ is $\binom{r}{j}$ therefore the left hand side of (1) is

$$\sum_{j=0}^{r} \binom{r}{j} (-1)^j$$

which is 0 if $0 < r$ and 1 if $r = 0$. □

Lemma 5 *If $G(\mathcal{C})$ has at least one edge then*

$$\sum_{\substack{\textit{matching of} \\ j \textit{ edges}}} (-1)^j + 2 \sum_{\textit{V-matching of}} (-1) + \sum_{\textit{N-matching}} (-1) \leq 0 \qquad (3)$$

where the matchings, V-matchings and N-matchings are subgraphs of $G(\mathcal{C})$.

Proof. The proof is analogous to the previous one. A mapping g can be defined from $\mathcal{M}_+$ to $\mathcal{M}_- \cup \mathcal{V} \cup \mathcal{N}$, basically in the same way as in the previous proof. □

Lemma 6

$$1+\sum_{j=1}^{\lfloor \frac{m}{2} \rfloor} \frac{(-1)^j}{j!} \binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2j+2}{2} 2^{-jH_2(\xi\to\eta)} -$$

$$-\sum_{j=0}^{\lfloor \frac{m-3}{2} \rfloor} \frac{1}{j!} 3 \binom{m}{3}\binom{m-3}{2}\binom{m-5}{2}\cdots\binom{m-2j-1}{2} p(\xi,\eta,V) 2^{-jH_2(\xi\to\eta)} -$$

$$-\sum_{j=0}^{\lfloor \frac{m-4}{2} \rfloor} \frac{1}{j!} 12 \binom{m}{4}\binom{m-4}{2}\binom{m-6}{2}\cdots\binom{m-2j-2}{2} p(\xi,\eta,N) 2^{-jH_2(\xi\to\eta)} \le$$

$$\le \Pr(\xi\to\eta, m) \le \tag{4}$$

$$\le 1+\sum_{j=1}^{\lfloor \frac{m}{2} \rfloor} \frac{(-1)^j}{j!} \binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2j+2}{2} 2^{-jH_2(\xi\to\eta)} +$$

$$+\sum_{j=0}^{\lfloor \frac{m-3}{2} \rfloor} \frac{1}{j!} 3 \binom{m}{3}\binom{m-3}{2}\binom{m-5}{2}\cdots\binom{m-2j-1}{2} p(\xi,\eta,V) 2^{-jH_2(\xi\to\eta)} +$$

$$+\sum_{j=0}^{\lfloor \frac{m-4}{2} \rfloor} \frac{1}{j!} 12 \binom{m}{4}\binom{m-4}{2}\binom{m-6}{2}\cdots\binom{m-2j-2}{2} p(\xi,\eta,N) 2^{-jH_2(\xi\to\eta)}.$$

Proof. The random pairs (ξ_i, η_i) $(1 \le i \le m)$ define a random partition on the set $\{1, 2, \ldots, m\}$ in a natural way, by the equality of these pairs: $C_{k,\ell} = \{i : (\xi_i, \eta_i) = (k, \ell)\}$. Then $C_k = \cup_\ell C_{k,\ell}$ is the kth class in the partition defined by ξ's. The event that η seems to be functionally dependent on ξ, that is, there is no pair $(k,\ell),(k,\ell')(\ell \neq \ell')$ among the m outcomes is equivalent to the event that $G(\mathcal{C})$ has no edge, that is, $\Pr(\xi \to \eta, m)$ equals $\Pr(G(\mathcal{C})$is an empty graph). In other words,

$$\Pr(\xi\to\eta, m) = \Pr(G(\mathcal{C})\text{is an empty graph}) +$$

$$\sum_{\mathcal{C}} 0 \cdot \Pr(\text{the pairs } (\xi_i, \eta_i) \text{ determine the partition } \mathcal{C}), \tag{5}$$

where the sum runs over all partitions with a non-empty $G(\mathcal{C})$. The elements $1, 2, \ldots, m$ are of course numbered, but the classes are not.

The left hand side of (2) is 1 for the $\mathcal{C}$ with the empty $G(\mathcal{C})$, otherwise it is non-negative by Lemma 4. Therefore, replacing the weights of the probabilities by this left hand side, an upper bound is obtained for (5):

$$\sum_{\mathcal{C}} \left(\sum_{\substack{\text{matching of} \\ j \text{ edges}}} (-1)^j + 2 \sum_{\text{V-matching}} 1 + \sum_{\text{N-matching}} 1 \right) \cdot$$

$$\Pr(\text{the pairs } (\xi_i, \eta_i) \text{ determine the partition } \mathcal{C}) \tag{6}$$

where the matchings, V-matchings and N-matchings are subgraphs of $G(\mathcal{C})$ for the given $\mathcal{C}$. Break this sum into 3 parts and consider first the part

$$\sum_{\mathcal{C}} \sum_{\substack{\text{matching of} \\ j \text{ edges}}} (-1)^j \Pr(\text{the pairs } (\xi_i, \eta_i) \text{ determine the partition } \mathcal{C}) =$$

$$\sum_{\substack{\text{matching of} \\ j \text{ edges}}} (-1)^j \sum_{\mathcal{C}} \Pr(\text{the pairs } (\xi_i, \eta_i) \text{ determine the partition } \mathcal{C}). \tag{7}$$

The last sum is nothing else but the probability of the event that all the edges in the given matching M are in $\mathcal{C}$, that is,

$$\Pr(\forall \{u, v\} \in M \text{ the relations } \xi_u = \xi_v, \eta_u \neq \eta_v \text{hold}).$$

Because of the independence, this is the jth power of $p(\xi, \eta, I)$ what is equal to

$$\sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2 = 2^{H_2(\xi \to \eta)} \tag{8}$$

by Lemma 1 and (1). We obtained

$$\sum_{\substack{\text{matching of} \\ j \text{ edges}}} (-1)^j 2^{-jH_2(\xi \to \eta)} \tag{9}$$

for (7).

The number of matchings consisting of j edges is

$$\frac{1}{j!}\binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2j+2}{2}.$$

Using this in (9), a new form of (7) is obtained:

$$1+\sum_{j=1}^{\lfloor\frac{m}{2}\rfloor}\frac{(-1)^j}{j!}\binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2j+2}{2}2^{-jH_2(\xi\to\eta)}$$

and this is the first row of the upper estimate in Lemma 6.

Now consider the second part of (6):

$$\sum_{\mathcal{C}}\sum_{\text{V-matching}}\Pr(\text{the pairs }(\xi_i,\eta_i)\text{ determine the partition }\mathcal{C})=$$

$$\sum_{\text{V-matching}}\sum_{\mathcal{C}}\Pr(\text{the pairs }(\xi_i,\eta_i)\text{ determine the partition }\mathcal{C}). \qquad (10)$$

The last sum is nothing else but the probability of the event that all the edges in the given V-matching V (containing $j+2$ edges) are in $\mathcal{C}$, that is,

$$\Pr(\forall\{u,v\}\in V\text{ the relations }\xi_u=\xi_v,\eta_u\neq\eta_v\text{ hold}).$$

Because of the independence, this is the jth power of (8) ($=2^{-H_2(\xi\to\eta)}$) times $p(\xi,\eta,V)$ what is given in Lemma 2. The result for (10) is

$$\sum_{\substack{\text{V-matching of}\\ j+2\text{ edges}}}\left(\sum_k p_k^3-2\sum_{k,\ell}p_kp_{k,\ell}^2+\sum_{k,\ell}p_{k,\ell}^3\right)2^{-jH_2(\xi\to\eta)}. \qquad (11)$$

Since the number of V-matchings is

$$\sum_{j=0}^{\lfloor\frac{m-3}{2}\rfloor}\frac{1}{j!}3\binom{m}{3}\binom{m-3}{2}\binom{m-5}{2}\cdots\binom{m-2j-1}{2},$$

(11) leads to a new form of (10), giving the second row of the upper estimate of Lemma 6.

The third row can be obtained in an analogous way, the only difference is that $p(\xi, \eta, N)$ should be used rather than $p(\xi, \eta, V)$. This finishes the proof of the upper bound.

The proof of the lower bound is the same, but Lemma 5 should be the starting point rather than Lemma 4. □

Lemma 7 *If*

$$2\log_2 m - H_2(\xi \to \eta) \to a \tag{12}$$

where a is a constant, independent on n and $m \to \infty$ then

$$1 + \sum_{j=1}^{\lfloor \frac{m}{2} \rfloor} \frac{(-1)^j}{j!} \binom{m}{2} \binom{m-2}{2} \cdots \binom{m-2j+2}{2} 2^{-jH_2(\xi \to \eta)} \tag{13}$$

tends to

$$e^{-2^{a-1}}.$$

Proof. The inequalities

$$\frac{(m-2j)^{2j}}{2^j} \le \binom{m}{2} \binom{m-2}{2} \cdots \binom{m-2j+2}{2} \le \frac{m^{2j}}{2^j}$$

lead to the following lower and upper estimates for (13):

$$-\sum_{j=1,3,\ldots,2j\le m} \frac{1}{j!} \cdot \frac{m^{2j}}{2^j} 2^{-jH_2(\xi\to\eta)} + 1 + \sum_{j=2,4,\ldots,2j\le m} \frac{1}{j!} \cdot \frac{(m-2j)^{2j}}{2^j} 2^{-jH_2(\xi\to\eta)} =$$

$$-\sum_{j=1,3,\ldots,2j\le m} \frac{1}{j!} 2^{j(2\log m - H_2(\xi\to\eta) - 1)} + 1 + \sum_{j=2,4,\ldots,2j\le m} \frac{1}{j!} 2^{j(2\log(m-2j) - H_2(\xi\to\eta) - 1)} \tag{14}$$

and

$$-\sum_{j=1,3,\ldots,2j\le m} \frac{1}{j!} \cdot \frac{(m-2j)^{2j}}{2^j} 2^{-jH_2(\xi\to\eta)} + 1 + \sum_{j=2,4,\ldots,2j\le m} \frac{1}{j!} \cdot \frac{(m)^{2j}}{2^j} 2^{-jH_2(\xi\to\eta)} =$$

$$-\sum_{j=1,3,\ldots,2j\le m} \frac{1}{j!} 2^{j(2\log(m-2j) - H_2(\xi\to\eta) - 1)} + 1 + \sum_{j=2,4,\ldots,2j\le m} \frac{1}{j!} 2^{j(2\log m - H_2(\xi\to\eta) - 1)} \tag{15}$$

Compare the members with $\log m$ and $\log(m-2j)$, respectively:

$$2^{j(2\log m - H_2(\xi\to\eta)-1)} - 2^{j(2\log(m-2j) - H_2(\xi\to\eta)-1)} = \tag{16}$$

$$2^{j(2\log m - H_2(\xi\to\eta)-1)}\left(1 - 2^{2j(\log(m-2j)-\log m)}\right) =$$

$$2^{j(2\log m - H_2(\xi\to\eta)-1)}\left(1 - \left(\frac{m-2j}{m}\right)^{2j}\right) =$$

$$2^{j(2\log m - H_2(\xi\to\eta)-1)}\left(1 - \left(1-\frac{2j}{m}\right)^{2j}\right).$$

Since $2j \le m$, the last factor can be upperbounded using the Bernoulli inequality:

$$1 - \left(1-\frac{2j}{m}\right)^{2j} \le 2j\frac{2j}{m} = \frac{4j^2}{m}.$$

Hence

$$2^{j(2\log m - H_2(\xi\to\eta)-1)}\frac{4j^2}{m} \tag{17}$$

is an upper bound for (16).

Consider the total change in (14) if the terms with $\log(m-2j)$ are replaced by terms with $\log m$ and use (17).

$$\sum_{j=2,4,\ldots,2j\le m}\frac{1}{j!}2^{j(2\log m - H_2(\xi\to\eta)-1)} - \sum_{j=2,4,\ldots,2j\le m}\frac{1}{j!}2^{j(2\log(m-2j) - H_2(\xi\to\eta)-1)} \le$$

$$\sum_{j=2,4,\ldots,2j\le m}\frac{1}{j!}2^{j(2\log m - H_2(\xi\to\eta)-1)}\frac{4j^2}{m}. \tag{18}$$

We need to show that this tends to 0 with n. Since $2\log m - H_2(\xi\to\eta) - 1$ tends to $a-1$, there is a threshold n_1 such that $2\log m - H_2(\xi\to\eta) - 1 \le a$ when $n > n_1$. Each term in (18) tends to 0, therefore the sum of the terms until $j \le n_1$ will do so. In the terms with $j > n_1$ the expression $2\log m - H_2(\xi\to\eta) - 1$ can be replaced by a without decreasing them. $\frac{1}{m}\frac{4j^2}{j!}2^{ja}$ is obtained as an upper bound for the jth term. Extend the sum with the odd terms and the large terms the following upper bound is obtained:

$$\frac{4}{m}\sum_{j=0}^{\infty}\frac{j^2}{j!}2^{ja} = \frac{4}{m}\left(2^{2(a+1)}e^{2^a} + 2^{a+1}e^{2^a}\right)$$

which obviously tends to 0 with $n \to \infty$. This shows that $\log(m-2j)$ can be replaced by $\log m$ in (14) without changing its limit for $n \to \infty$. Then (14) becomes

$$e^{-2^{2\log m - H_2(\xi\to\eta)-1}}$$

which tends to $e^{-2^{a-1}}$. Therefore the lim inf of (13) is at least this much. Starting from (15), the same steps prove that that the lim sup of (13) cannot be more. This is really its limit. □

Lemma 8 *Suppose that $m \to \infty$, (12) and*

$$\frac{p(\xi,\eta,V)^2}{p(\xi,\eta,I)^3} \to 0 \tag{19}$$

hold. Then

$$\sum_{j=0}^{\lfloor\frac{m-3}{2}\rfloor} \frac{1}{j!} 3\binom{m}{3}\binom{m-3}{2}\binom{m-5}{2}\cdots\binom{m-2j-1}{2} p(\xi,\eta,V) 2^{-jH_2(\xi\to\eta)} \to 0.$$

Proof. It will be similar to that of Lemma 7. Start with the upper estimate

$$\binom{m-3}{2}\binom{m-5}{2}\cdots\binom{m-2j-1}{2} \le \frac{m^{2j}}{2^j}.$$

This leads to the following upper estimate for the investigated quantity:

$$3\binom{m}{3} p(\xi,\eta,\mathrm{V}) \sum_{j=0}^{\infty} \frac{m^{2j}}{2^j j!} 2^{-jH_2(\xi\to\eta)} =$$

$$3\binom{m}{3} p(\xi,\eta,\mathrm{V}) \sum_{j=0}^{\infty} \frac{1}{j!} 2^{j(2\log m - H_2(\xi\to\eta)-1)} =$$

$$3\binom{m}{3} p(\xi,\eta,\mathrm{V}) e^{2^{2\log m - H_2(\xi\to\eta)-1}}.$$

Here the last factor tends to $e^{2^{a-1}}$ by (12), therefore we only have to show that

$$3\binom{m}{3} p(\xi,\eta,\mathrm{V}) \to 0. \tag{20}$$

(12) implies

$$m^2\left(\sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2\right) \to 2^a$$

and

$$m^3\left(\sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2\right)^{\frac{3}{2}} \to 2^{\frac{3a}{2}}.$$

This convergence and the square root of (19) prove (20). □

Lemma 9 *Suppose that $m \to \infty$, (12) and*

$$\frac{p(\xi,\eta,N)}{p(\xi,\eta,I)^2} \to 0 \tag{21}$$

hold. Then

$$\sum_{j=0}^{\lfloor\frac{m-4}{2}\rfloor} \frac{1}{j!} 12 \binom{m}{4}\binom{m-4}{2}\binom{m-6}{2}\cdots\binom{m-2j-2}{2} p(\xi,\eta,N) 2^{-jH_2(\xi\to\eta)} \to 0.$$

Proof. It is analogous to the previous one. □

Now the statement of the theorem is an easy consequence of Lemmas 6-9. □

4 Previous work, remarks, future work

Related earlier work. The problem in question was studied in the papers of Selivanov [8] and Mihailov and Selivanov [3]. They have proved limit theorems on the convergence of the quantity studied here to the Poisson and normal distributions, respectively.

Our motivation: database theory. Our primary motivation was database theory. A very simple model of a database is an $m \times n$ matrix, where the columns are representing the types of data (called *attributes*), say last name, first name, etc. while the data of one individual are in one row. A fundamental concept in the theory is the *functional dependency*. Let A be a set of columns, b one column. We say that b *functionally depends* on A if the individuals having the same data in the columns belonging to A have

the same data in b. Shortly, the data in A uniquely determine the data in b. More precisely, the matrix has no two rows having the same entries in the columns in A and different in b. In notation $A \to b$. In most of the older works it is supposed that there is a "logical connection" among the data, so the functional dependencies are a priori given. Here we adopt the view that only those functional dependencies $A \to b$ exist which are determined by the given matrix.

Suppose that some probabilistic connections are a priori given among the data, that is a joint distribution

$$\Pr(\zeta_1 = u_1, \zeta_2 = u_2, \dots, \zeta_n = u_n)$$

is given among the n data in one row. (We might know or we might not know this distribution.) The choice of the rows is totally independent. Let ζ_A be the random vector with the components ζ_i for all $i \in A$. Of course, the distribution of a row determines the joint distribution of the pair ζ_A, ζ_b. For fixed n, A and b we could speak about the probability $\Pr(\zeta_A \to \zeta_b, m)$ of the event that the m actual rows indicate that $A \to b$. This situation leads to the problem only mentioned in the Introduction, but not considered in the present paper.

Now we describe our real motivating problem. Suppose that n is large, the m (it is a function of n) rows of the matrix are chosen following the given joint distribution. What are the sizes of A satisfying $A \to b$ for some b, that is, what are the typical sizes of the functional dependencies appearing in the matrix. It is intuitively clear that for small (say of fixed size) A this cannot happen (unless the distribution gives a functional dependency). The sizes of the A's showing $A \to b$ must increase by n. Then ζ_A as a vector of growing size has an increasing number of possible values, and their probabilities are typically decreasing. This is how we arrived to the model of the paper when the probabilities of ξ are decreasing with n. We will show in a forthcoming paper how to use the results of the present paper for the determination of the typical sizes of A's in a functional dependency $A \to b$ in a large database.

The special case when the ζ_i's are independent was considered in [2]. The method of the present paper is a generalization of that paper. Similar (but not identical) results using different methods can be found in [1]. Papers [6] and [7] contain somewhat related results on random databases.

On the conditions of the main theorem. The two conditions ((19) and (21)) are chosen by a very simple reason: the proof works under them.

When are they satisfied? It is easy to see that if the probabilities "uniformly" tend to 0, ξ and η are "nearly independent" that is there are constants c, d, C, D such that

$$\frac{C}{n} < p_k < \frac{D}{n}, \qquad \frac{c}{n^2} < p_{k,\ell} < \frac{d}{n^2}$$

hold then (19) and (21) are satisfied. On the other hand, if one p_k does not tend to 0 then the conditions are not satisfied. More work is needed to find necessary and sufficient conditions for the probability distributions under which these conditions are true. We do not even know whether the two conditions are independent or not. Does (19) imply (21)?

Our function $H_2(\xi \to \eta)$, special cases. It is slightly related to the Rényi entropy of order 2 (see [4] and [5]):

$$H_2(\xi) = -\log_2 \sum_k p_k^2.$$

However our formula (1) is far from being a "conditional entropy" derived from the Rényi entropy.

If η is a function of ξ then there is a unique ℓ for which $p_{k,\ell}$ is non-zero, that is, $p_{k,\ell} - p_k$. Hence $p(\xi, \eta, I) = \sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2 = 0$ and $H_2(\xi \to \eta) = \infty$. The trivial statement $\Pr(\xi \to \eta, m) = 1$ in this case really follows from Theorem 1.

Suppose now that ξ and η are independent. Define $q_\ell = \sum_k p_{k,\ell}$. Also suppose that η is not "nearly one-valued" that is there is no ℓ for which q_ℓ is near to 1 for infinitely many n. More precisely we suppose that there is an ε such that $1 - \varepsilon > \sum_\ell q_\ell^2$ for large n's. Then

$$\sum_{k,\ell} p_{k,\ell}^2 = \sum_{k,\ell} p_k^2 q_\ell^2 = \sum_k p_k^2 \sum_\ell q_\ell^2$$

therefore

$$p(\xi, \eta, I) = \sum_k p_k^2 - \sum_{k,\ell} p_{k,\ell}^2 = \sum_k p_k^2 \left(1 - \sum_\ell q_\ell^2\right)$$

and

$$H_2(\xi \to \eta) = -\log_2 \left(\sum_k p_k^2\right) - \log_2 \left(1 - \sum_\ell q_\ell^2\right)$$

hold. The second term on the right hand side is upperbounded by $\log_2 \varepsilon$, while the first term tends to infinity by (19). Hence $H_2(\xi \to \eta)$ asymptotically depends only on ξ. By Theorem 1, the same is implied for $\Pr(\xi \to \eta, m)$ as it is expected in this case.

5 Acknowledgement

We are indebted to the anonymous referee for pointing out the weaknesses of the first version of the paper.

References

[1] Demetrovics, J., Katona, G.O.H., Miklós, D., Seleznjev, O., Thalheim, B., Asymptotic properties of keys and functional dependencies in random databases, *Theoretical Computer Sciences* **190**(1998) 151-166.

[2] Demetrovics, J., Katona, G.O.H., Miklós, D., Seleznjev, O., Thalheim, B., Functional dependencies in random databases, *Studia Sci. Math. Hungar.* **34**(1998) 127-140.

[3] Mikhailov, V.G., Selivanov, B.I. On some statistics in a scheme of particles allocation in cells of a two-dimensional table, in: *Tr. Diskr. Mat.*, **4**, Fizmatlit, Moscow, 2001, pp. 169176

[4] Rényi Alfréd, Some fundamental questions of information theory (in Hungarian), *MTA III Oszt. Közl.* **10**(1960) 251-282.

[5] Rényi Alfréd, On measures of information and entropy, *Proc. of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, 1960*, 1961, pp. 547-561.

[6] Seleznjev, Oleg, Thalheim, Bernhard, Average Case Analysis in Database Problems, *Methodology and Computing in Applied Probability* **5**(4)(2003) 395-418.

[7] Seleznjev, Oleg, Thalheim, Bernhard, Random Databases with Approximate Record Matching, *Methodology and Computing in Applied Probability* **12**(1)(2010) 63-89.

[8] Selivanov, B.I., On a generalization of the classical allocation problem, *Teor. Veroyatnost. i Primenen.* **43**(2)(1998), 315330