# So Far Yet So Close: How Independence is Almost Dependence

Himadri Mandal

## Introduction

Imagine two continuous random variables seemingly unrelated and free from any influence from each other. It's a common assumption in probability theory that independence implies a lack of any relationship between variables.

However, a fascinating theorem challenges this notion, showing that from any two independent continuous random variables, we can construct new deterministically dependent variables that not only share the same distributions but also have joint distributions that are arbitrarily close.

In other words, independence can be deceptive, leading us to a realm where variables are closer than we think. Let's explore this intriguing theorem and its implications.

This theorem was proven by Roger B. Nelsen in his book "An introduction to Copulas", theorem 3.2.2. Here I will be sharing the the proof my friend Himadri Mandal came up with.

### Theorem

Let $X, Y$ be independent continuous random variables on $(\mathbb{R}, \mathcal{B}, P)$. For all $\epsilon > 0$ there exist deterministically dependent random variables $U, V$ on the same probability space such that $U \sim X,\ V \sim Y$ and

$$\sup_{(a,b)\in\mathbb{R}^2} |F_{U,V}(a,b) - F_{X,Y}(a,b)| < \epsilon$$

.

## Proof

We will prove this for continuous random variables with range [0,1]. For instance if $X$ is a random variable with range R, we consdier $\frac{1}{1+e^X}$ instead, to bound it in [0,1]. Now fix n $\in \mathbb{N}$ and obtain two partitions of [0,1] in the following manner.

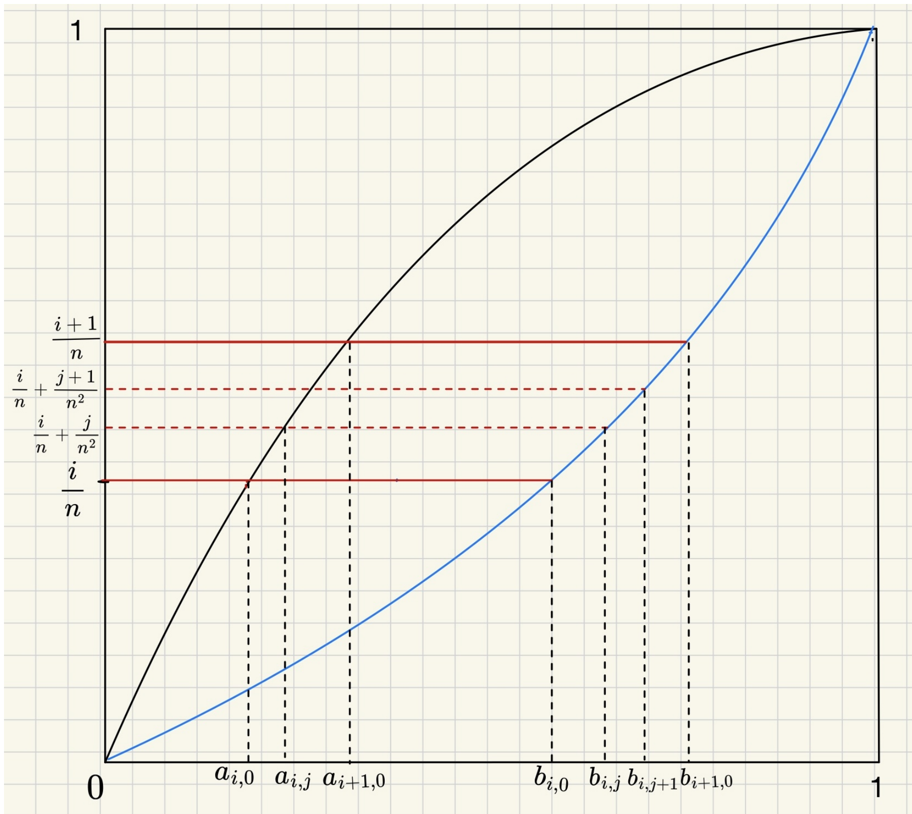


Figure 1: Partitioning CDFs

First consider the CDFs of X and Y. That is bounded in [0,1]. Make $n^2$ many partitions of [0,1] in the Y-axis each having length $\frac{1}{n^2}$ each. Now consider the corresponding intervals in the X-axis. For the CDF of X call $a_{ij} = F_X^{-1}(\frac{i}{n} + \frac{j}{n^2})$. Similarly define $b_{ji}$ for Y. Since CDFs are bijective continuous functions, we obtain

$$\{a_{0,1} = 0 < \cdots < a_{0,n-1} < a_{1,0} < \cdots < a_{1,n-1} < \cdots < a_{n-1,n} = 1\}$$

$$\{b_{0,1} = 0 < \cdots < b_{0,n-1} < b_{1,0} < \cdots < b_{1,n-1} < \cdots < b_{n-1,n} = 1\}$$

Clearly,

$$F_X(a_{j,0}) - F_X(a_{j-1,0}) = \frac{1}{n},\ F_Y(b_{j,0}) - F_Y(b_{j-1,0}) = \frac{1}{n}$$

$$F_X(a_{j,i}) - F_X(a_{j,i-1}) = \frac{1}{n^2},\ F_Y(b_{j,i}) - F_Y(b_{j,i-1}) = \frac{1}{n^2}$$

Define,

$$I_{i,j} := \{a_{i,j} \leq x \leq a_{i,j+1}\},\ J_{j,i} := \{b_{j,i} \leq y \leq b_{j,i+1}\}$$

$$I_i = \cup_j I_{i,j},\ \ J_j = \cup_i J_{j,i}$$

Now our aim is to generate a random variable $g_n(Y) = U_n$ where $g_n : [0,1] \to [0,1]$ is a bijective function such that $g_n(Y) \sim X$ and

$$U_n \in I_{i,j} \iff Y \in J_{j,i} \cdots (1)$$

To determine $g_n$ we back-trace it using it's desirable properties.
$g_n(Y) \sim X \iff Y \sim g_n^{-1}(X)$. And hence for $y \in J_{j,i}$ we need

$$F_Y(y) - F_Y(b_{j,i}) = P\left(b_{j,i} \le Y \le y\right) \overset{!}{=} P\left(b_{j,i} \le g_n^{-1}(X) \le y\right)$$

$$= P\left(g_n(b_{j,i}) \le X \le g_n(y)\right) \overset{!}{==} P\left(a_{i,j} \le X \le g_n(y)\right)\ [We\ want\ g_n(b_{j,i}) = a_{i,j}]$$

$$= F_X(g_n(y)) - F_X(a_{i,j})$$

$$Hence,\ F_X(g_n(y)) \overset{!}{=} F_Y(y) + (F_X(a_{i,j}) - F_Y(b_{j,i}))$$

$$= F_Y(y) + \left((i-j) \cdot \left(\frac{n-1}{n^2}\right)\right) \cdots (2)$$

Thus we get for $y \in J_{j,i}$

$$\boxed{g_n(y) \overset{!}{=} F_X^{-1}\left(F_Y(y) + \left((i-j) \cdot \left(\frac{n-1}{n^2}\right)\right)\right)}$$

This is clearly measurable, and this function indeed satisfies condition (1) [CHECK!] Now we show that $g_n(Y) \sim X$. For $x \in I_{i,j}$

$$P\left(a_{i,j} \le U \le x\right)$$

$$= P\left(a_{i,j} \le F_X^{-1}\left(F_Y(y) + \left((i-j) \cdot \left(\frac{n-1}{n^2}\right)\right)\right) \le x\right)$$

$$= P\left(F_X(a_{i,j}) \le F_Y(y) + \left((i-j) \cdot \left(\frac{n-1}{n^2}\right)\right) \le F_X(x)\right) ......(\star)$$

$$= F_X(x) - F_X(a_{i,j}) = P(a_{i,j} \le X \le x)......(\star\star)$$

But why is the last equality true? First consider

$$P\left(F_X(a_{i,j}) \le F_Y(y) \le F_X(x)\right)$$

This is nothing but $F_X(x) - F_X(a_{i,j})$. Observe that

$$P(F_X(a_{i,j}) \le F_Y(y) + C \le F_X(x)) = F_X(x) - F_X(a_{i,j})$$

where C is a constant and $C \le F_X(a_{i,j})$. In $(\star)$ the constant term is

$$(i-j) \cdot \left(\frac{n-1}{n^2}\right) = F_X(a_{i,j}) - F_Y(b_{j,i})$$

$$F_X(a_{i,j}) - (F_X(a_{i,j}) - F_Y(b_{j,i})) = F_Y(b_{j,i}) \ge 0$$

Hence $(\star\star)$ holds. This proves $g_n(Y) \sim X$.

## Lemma 1

$$P\left(X \in I_i\right) \cdot P\left(Y \in J_j\right) = P\left(X \in I_i \text{ and } Y \in J_j\right)$$
$$= P\left(U_n \in I_i \text{ and } Y \in J_j\right) \ \forall\ i, j$$

## Proof

By the independence of $X, Y$ ,

$$P\left(X \in I_i\right) \cdot P\left(Y \in J_j\right) = P\left(X \in I_i,\ Y \in J_j\right) = \frac{1}{n^2}.$$

RHS=

$$P\left(U \in \cup_j I_{i,j},\ Y \in J_j\right) = \sum_{k=0}^{n-1} P\left(U \in I_{i,k}, Y \in J_j\right) = P(U \in I_{i,j},\ Y \in J_j)$$

$$= \sum_{l=0}^{n-1} P(U \in I_{i,j},\ Y \in J_{j,l}) = P(U \in I_{i,j},\ Y \in J_{j,i}) = P(U \in I_{i,j})[as\ (1)\ holds]$$

$$= P(X \in I_{i,j}) = \frac{1}{n^2} \quad \square$$

## Lemma 2

$$|F_{U_n,Y}(a,b) - F_{X,Y}(a,b)| < \frac{4}{n} \qquad \forall a, b \in [0,1]$$

## Proof

At first we disjointify a and b in the following manner.
Let $(0,a) = \{\cup_{i=0}^{m-1} I_i\} \cup A$ and $(0,b) = \{\cup_{i=0}^{k-1} J_i\} \cup B$ where $a \in I_m and\ b \in J_k$.
$A \subset I_m$ and $B \subset J_k$ are the residual intervals. Hence

$$P(X \in A) < \frac{1}{n};\ P(Y \in B) < \frac{1}{n}$$

We disjointify [0,1]× [0,1] to get

$$P(X \leq a,\ Y \leq b) = \sum_{i=0}^{m-1} \sum_{j=0}^{k-1} P(X \in I_i,\ Y \in J_j) + P(X \in A, Y \leq b)$$
$$+ P(X \leq a_{m,0},\ Y \in B)$$
$$P(U_n \leq a,\ Y \leq b) = \sum_{i=0}^{m-1} \sum_{j=0}^{k-1} P(U_n \in I_i,\ Y \in J_j) + P(U_n \in A, Y \leq b)$$
$$+ P(U_n \leq a_{m,0},\ Y \in B)$$

As an implication of Lemma 1 we get,

$$\sum_{i=0}^{m-1}\sum_{j=0}^{k-1} P(X \in I_i,\ Y \in J_j) = \sum_{i=0}^{m-1}\sum_{j=0}^{k-1} P(U \in I_i,\ Y \in J_j)$$

Hence,

$$|F_{U_n,Y}(a,b) - F_{X,Y}(a,b)| < P(X \in A) + P(Y \in B) + P(U_n \in A) + P(Y \in B)$$

$$< 2(P(X \in A) + P(Y \in B)) < \frac{4}{n} \quad \square$$

Now as $n \to \infty$ the joint distributions get arbitrarily close. This completes the proof of the theorem. Here $U_n, Y$ serves as our required $U, V$. ■

**Note:**The inequality proved here is weak and can me made much stronger, however for proving the theorem this inequality was enough to prove.