INTEGERS: ELECTRONIC JOURNAL OF COMBINATORIAL NUMBER THEORY 6 (2006), #A04

SIMPSON'S PARADOX IN THE FAREY SEQUENCE

Rasa Šleževičienė-Steuding
Departamento de Matemáticas, Universidad Autónoma de Madrid, C. Universitaria de Cantoblanco, 28 049 Madrid, Spain
rasa.steuding@uam.es

Jörn Steuding
Departamento de Matemáticas, Universidad Autónoma de Madrid, C. Universitaria de Cantoblanco, 28 049 Madrid, Spain
jorn.steuding@uam.es

Received: 6/6/05, Revised: 12/2/05, Accepted: 2/4/06, Published: 2/14/06

Abstract

We investigate the appearance of Simpson's paradox in the Farey sequence of reduced fractions in the unit interval.

1. Introduction and statement of results

In statistics it frequently occurs that the data seems to contradict our intuition. For instance, Cohen & Nagel [1] cited actual death rates from tuberculosis in Richmond (Virginia) and New York from 1910 that verified the following propositions:

- the death rate for African Americans was lower in Richmond than in New York,
- the death rate for Caucasians was lower in Richmond than in New York,
- the death rate for the total combined population of African Americans and Caucasians was higher in Richmond than in New York.

How can it be? We may illustrate this with another numerical example, namely

$$\frac{3}{5} < \frac{8}{13} \qquad \text{and} \qquad \frac{7}{10} < \frac{5}{7}, \qquad \text{but} \qquad \frac{3+7}{5+10} = \frac{2}{3} > \frac{13}{20} = \frac{8+5}{13+7}.$$

This phenomenon is called Simpson's paradox after E.H. Simpson [6] who published in 1951 an influential paper on this topic. Of course, Simpson's paradox is not a paradox but just a

simple fact about fractions. Nevertheless, it has a variety of surprising applications arising from links between proportions, probabilities, and their representations as fractions. A short historical overview on Simpson's paradox can be found in Mittal [5].

Usually, statistical data is given as some subset of the set of rational numbers $\mathbb{Q}$. In this note we investigate the appearance of Simpson's paradox within $\mathbb{Q}$. For an enumeration of $\mathbb{Q}$ we shall use the Farey sequence of reduced fractions in the unit interval.

For $n \in \mathbb{N}$ the Farey sequence $\mathcal{F}_n$ is the ordered list of all reduced fractions in the unit interval having denominators less than or equal to n, i.e.,

$$\mathcal{F}_n := \left\{ \frac{a}{b} \in \mathbb{Q} \,:\, 0 \le a \le b \le n,\, \gcd(a,b) = 1 \right\},$$

where, as usual, $\gcd(a,b)$ denotes the greatest common divisor of the integers a and b. Consecutive Farey fractions $\frac{a}{b} < \frac{c}{d}$ satisfy $bc - ad = 1$, resp.

$$\frac{c}{d} - \frac{a}{b} = \frac{1}{bd}.$$

Thus, $\mathcal{F}_n$ is not equidistantly distributed; this property makes Farey fractions useful tools in the theory of Diophantine approximations (see the classical paper of Ford [2]). The Farey sequence can be build from $\mathcal{F}_1$ by taking mediants of $\frac{0}{1}$ and $\frac{1}{1}$. For $\frac{a}{b}, \frac{c}{d} \in \mathcal{F}_n$ their mediant is defined by $\frac{a+c}{b+d}$. If $\frac{a}{b} \le \frac{c}{d}$, then it is easily seen that

$$\frac{a}{b} \le \frac{a+c}{b+d} \le \frac{c}{d};$$

equality holds if and only if $\frac{a}{b} = \frac{c}{d}$. For consecutive elements $\frac{a}{b}, \frac{c}{d} \in \mathcal{F}_n$, their mediant is an element of $\mathcal{F}_{b+d}$. The limit of the Farey sequence, $\cup_{n\in\mathbb{N}}\mathcal{F}_n$, consists exactly of the reduced fractions in the interval $[0,1]$.

It is a natural question to ask *how often* Simpson's paradox occurs in the Farey sequence. So we are interested in the system of diophantine inequalities

$$0 \le \frac{a}{b} < \frac{A}{B}, \quad 0 \le \frac{c}{d} < \frac{C}{D}, \qquad \text{and} \qquad \frac{a+c}{b+d} > \frac{A+C}{B+D} \tag{1}$$

for given integers A, B, C, D. It is easily seen that a certain asymmetry is necessary for this phenomenon to happen. Assume that $BC = AD$, resp. $\frac{A}{B} = \frac{C}{D}$. Then, obviously, the mediant $\frac{A+C}{B+D}$ coincides with both, which shows that $\frac{a}{b}$ and $\frac{c}{d}$ both are less than $\frac{A+C}{B+D}$. Further, the case of $A = 0$ is uninteresting. Therefore, in the sequel we may assume w.l.o.g. that $0 < \frac{A}{B} < \frac{C}{D} \le 1$.

First of all, we note that there are infinitely many examples for Simpson's paradox everywhere in the Farey sequence. As a matter of fact, it is elementary to check that for any positive integer n with $\gcd(n,B) = 1$ and $\gcd(n,D) = 1$, and satisfying the inequalities:

$$0 < \frac{a}{b} = \frac{A}{B} - \frac{1}{n} < \frac{A}{B} \qquad \text{and} \qquad 0 < \frac{c}{d} = \frac{C}{D} - \frac{1}{n^2} < \frac{C}{D},$$

then

$$\frac{a+c}{b+d} > \frac{A+B}{C+D} \qquad \text{for} \quad n > \frac{1}{2}\left(1 + \sqrt{1 + 4\frac{(B+D)^2}{BC-AD}}\right).$$

The last example gives a slight indication as to whether or not Simpson's paradox is a rare event. We shall prove that, given two fractions $0 < \frac{A}{B} < \frac{C}{D}$ in the Farey sequence $\mathcal{F}_n$, pairs $(\frac{a}{b}, \frac{c}{d}) \in \mathcal{F}_n^2$ satisfying (1) occur with positive probability (in the sense of a Laplace experiment), as $n \to \infty$.

Theorem. *For two fractions $0 < \frac{A}{B} < \frac{C}{D} \le 1$ in the Farey sequence $\mathcal{F}_n$, we have*

$$\lim_{n\to\infty} \frac{\sharp\left\{\left(\frac{a}{b}, \frac{c}{d}\right) \in \mathcal{F}_n^2 \,:\, \frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D} \quad \text{and} \quad \frac{a+c}{b+d} > \frac{A+C}{B+D}\right\}}{\sharp\left\{\left(\frac{a}{b}, \frac{c}{d}\right) \in \mathcal{F}_n^2 \,:\, \frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D}\right\}} = \delta > 0$$

with

$$\delta := \delta\left(\frac{A}{B}, \frac{C}{D}\right) := \frac{1}{9}\frac{BD}{AC}\left\{\frac{1}{2}\Delta^2\Big(D^2\Psi(1) - 2BD\Psi(0) + B^2\Psi(-1)\Big) + \right.$$
$$\left. + \frac{A}{B}\Big(B\Delta\Upsilon(0) - \Big(D\Delta + \frac{1}{2}\frac{A}{B}\Big)\Upsilon(1)\Big)\right\},$$

where

$$\Upsilon(\ell) := \frac{36}{4-\ell^2}\min\left\{1, \nabla^{-2-\ell}\right\} - \frac{9}{2-\ell}\min\left\{\nabla^{2-\ell}, \nabla^{-2-\ell}\right\}, \tag{2}$$

$$\Psi(\ell) := \frac{9}{2-\ell}\Big(\min\left\{\nabla^{2-\ell}, \nabla^{-2-\ell}\right\} - \min\left\{\left(\frac{D}{B}\right)^{2-\ell}, \left(\frac{D}{B}\right)^{-2-\ell}\right\}\Big) + \tag{3}$$
$$+ \frac{36}{4-\ell^2}\Big(\min\left\{1, \left(\frac{D}{B}\right)^{-2-\ell}\right\} - \min\left\{1, \nabla^{-2-\ell}\right\}\Big),$$

$$\Delta := \frac{BC-AD}{BD(B+D)} \qquad \text{and} \qquad \nabla := \frac{D(A+C)}{BC-AD}. \tag{4}$$

The expression δ looks rather complicated on first sight but it simply is a rational function of A, B, C, D.

Note that there also exist examples with equality between the mediants. For instance,

$$\frac{1}{5} < \frac{2}{5}, \quad \frac{14}{25} < \frac{3}{5}, \qquad \text{and} \qquad \frac{1+14}{5+25} = \frac{1}{2} = \frac{2+3}{5+5}.$$

As the proof of the theorem will show the set of such examples has zero-density.

2. Proof of the theorem

There is an interesting geometrical interpretation of the Farey sequence. The Farey fractions $\frac{a}{b} \in \mathcal{F}_n$ correspond to the points with integer coordinates (a, b) situated in the triangle given by $\{(x, y) \in \mathbb{R}^2 \,:\, 0 \le x, y, x + y \ge n\}$, which *can be seen from the origin*, i.e., for which a and b are coprime. As we shall see below, for the proof of the theorem we have to count lattice points under this and further conditions.

We start with asymptotic formula for the number of Farey fractions $\frac{a}{b} \in \mathcal{F}_n$ under a given magnitude. Let $\xi \in (0, 1]$ be fixed, then

$$\sharp\left\{\frac{a}{b} \in \mathcal{F}_n \,:\, \frac{a}{b} \le \xi\right\} = 1 + \sum_{1\le b\le n} \sum_{\substack{1\le a\le b\xi \\ \gcd(a,b)=1}} 1 = \Sigma(n;\xi),$$

say. Following the argument which gives an asymptotic formula for the cardinality of $\mathcal{F}_n$ (see [3]), we find

$$\Sigma(n;\xi) = \frac{3\xi}{\pi^2} n^2 + O(n\log n), \tag{5}$$

where the implicit constant in the error term is absolute. Thus

$$\begin{aligned} \sharp\left\{\left(\frac{a}{b}, \frac{c}{d}\right) \in \mathcal{F}_n^2 \,:\, \frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D}\right\} &= \Sigma\left(n; \frac{A}{B}\right)\Sigma\left(n; \frac{C}{D}\right) \\ &= \frac{9}{\pi^4}\frac{AC}{BD} n^4 + O(n^3 \log n). \end{aligned} \tag{6}$$

In order to estimate the proportion in question we have to do a similar computation under the additional restriction (1) for the related mediants. Let

$$\sharp\left\{\left(\frac{a}{b}, \frac{c}{d}\right) \in \mathcal{F}_n^2 \,:\, \frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D} \quad \text{and} \quad \frac{a+c}{b+d} > \frac{A+C}{B+D}\right\} = \Sigma(n) + \Sigma_0(n),$$

where $\Sigma_0(n)$ counts the number of those tuples $(\frac{a}{b}, \frac{c}{d})$ for which $ac = 0$. It follows from (5) that $\Sigma_0(n) \ll n^2$ and so their contribution is negligible. We have

$$\Sigma(n) = \sum_{1\le b\le n} \sum_{1\le d\le n} \sum_{1\le a<b\frac{A}{B}, \gcd(a,b)=1} \sum_{\substack{1\le c<d\frac{C}{D}, \gcd(c,d)=1 \\ a+c>(b+d)\frac{A+C}{B+D}}} 1. \tag{7}$$

Denote by $\mu(n)$ the Möbius μ-function, i.e., $\mu(1) = 1$, $\mu(n) = (-1)^\nu$ if n is the product of ν distinct primes, and $\mu(n) = 0$ otherwise (if n has some square divisor). In view of the well-known formula

$$\sum_{d|m} \mu(d) = \begin{cases} 1 & \text{if} \quad m = 1, \\ 0 & \text{otherwise,} \end{cases}$$

we can rewrite (7) as

$$\Sigma(n) = \sum_{1\le b\le n} \sum_{1\le d\le n} \sum_{1\le a<b\frac{A}{B}} \sum_{\substack{1\le c<d\frac{C}{D} \\ a+c>(b+d)\frac{A+C}{B+D}}} \sum_{\alpha|\gcd(a,b)} \mu(\alpha) \sum_{\gamma|\gcd(c,d)} \mu(\gamma)$$

$$= \sum_{1\le b\le n}\sum_{1\le d\le n}\sum_{\alpha|b}\mu(\alpha)\sum_{\gamma|d}\mu(\gamma)\sum_{\substack{1\le a<b\frac{A}{B}\\ \alpha|a}}\sum_{\substack{1\le c<d\frac{C}{D}\\ \gamma|c,a+c>(b+d)\frac{A+C}{B+D}}} 1.$$

First we shall consider the inner two sums, which count the lattice points $(x,y)\in\mathbb{Z}^2$ with $x\equiv 0 \bmod \alpha, y\equiv 0 \bmod \gamma$ lying above the straight line $x+y=(b+d)\frac{A+C}{B+D}$ inside the rectangle $1\le x<b\frac{A}{B}, 1\le y<d\frac{C}{D}$. Since these are exactly the lattice points of the sublattice $\alpha\mathbb{Z}\times\gamma\mathbb{Z}$ in the convex region

$$\mathcal{R}(b,d):=\left\{(x,y)\in\mathbb{R}^2 : 1\le x<b\frac{A}{B}, 1\le y<d\frac{C}{D}, x+y>(b+d)\frac{A+C}{B+D}\right\},$$

the number of these lattice points equals asymptotically the volume of $\mathcal{R}(b,d)$ divided by the volume of a fundamental parallelepiped. It is not difficult to see that the error term is of the order of the boundary (see Lemma 2.1.1 in Huxley [4]). Thus,

$$\sum_{\substack{1\le a<b\frac{A}{B}\\ \alpha|a}}\sum_{\substack{1\le c<d\frac{C}{D}\\ \gamma|c,a+c>(b+d)\frac{A+C}{B+D}}} 1 = \frac{\mathsf{vol}(\mathcal{R}(b,d))}{\alpha\gamma}+O\left(\frac{\mathsf{length}(\mathcal{R}(b,d))}{\alpha+\gamma}\right). \tag{8}$$

Note that the volume of $\mathcal{R}(b,d)$ does not depend on α and γ. In the next step we shall compute this volume.

By simple geometric arguments it follows that the region $\mathcal{R}(b,d)$ is

- a single point or empty if and only if
$$b\frac{A}{B}+d\frac{C}{D}\le(b+d)\frac{A+C}{B+D};$$
- a triangle of volume
$$\frac{1}{2}\Delta^2(dB-bD)^2$$
(where Δ is defined by (4)) if and only if
$$d\frac{C}{D}\le(b+d)\frac{A+C}{B+D}<b\frac{A}{B}+d\frac{C}{D};$$
- a trapezoid of volume
$$\frac{A}{B}\left(bdB\Delta-b^2\left(D\Delta+\frac{1}{2}\frac{A}{B}\right)\right)$$
if and only if
$$(b+d)\frac{A+C}{B+D}<d\frac{C}{D}.$$

Other cases of intersections of a half-plane with a rectangle cannot occur; to see this note that $b < b+d$ and $\frac{A}{B} < \frac{A+B}{C+D}$, and thus

$$b\frac{A}{B} < (b+d)\frac{A+C}{B+D}.$$

Let us remark that all of the above listed cases occur for any given $0 < \frac{A}{B} < \frac{C}{D}$ as $n \to \infty$. In fact, this will imply, as we shall see below, the existence and the positivity of the proportion $\delta = \delta(\frac{A}{B}, \frac{C}{D})$.

The contribution of the main term from (8) to $\Sigma(n)$ may be written as

$$\Sigma(n) = \Sigma_{\Delta}(n) + \Sigma_{\diamondsuit}(n) + \mathsf{error}(n), \tag{9}$$

where $\Sigma_{\Delta}(n)$ and $\Sigma_{\diamondsuit}(n)$ are the sums of the terms with $\mathcal{R}(b,d)$ triangle or trapezoid, i.e.,

$$\begin{aligned}
\Sigma_{\Delta}(n) &:= \sum_{1\le b\le n} \sum_{\substack{1\le d\le n \\ d\frac{C}{D} \le (b+d)\frac{A+C}{B+D} < b\frac{A}{B} + d\frac{C}{D}}} \sum_{\alpha|b} \frac{\mu(\alpha)}{\alpha} \sum_{\gamma|d} \frac{\mu(\gamma)}{\gamma} \cdot \frac{1}{2}\Delta^2 (bD - dB)^2, \\
\Sigma_{\diamondsuit}(n) &:= \sum_{1\le b\le n} \sum_{\substack{1\le d\le n \\ (b+d)\frac{A+C}{B+D} < d\frac{C}{D}}} \sum_{\alpha|b} \frac{\mu(\alpha)}{\alpha} \sum_{\gamma|d} \frac{\mu(\gamma)}{\gamma} \cdot \frac{A}{B}\left(bdB\Delta - b^2\left(D\Delta + \frac{1}{2}\frac{A}{B}\right)\right),
\end{aligned}$$

and $\mathsf{error}(n)$ is the error term with respect to (8). We may rewrite the condition of summation for the triangle and for the trapezoid by

$$b\frac{D}{B} < d \le b\frac{D(A+C)}{BC-AD} = b\nabla \qquad \text{and} \qquad b\nabla = b\frac{D(A+C)}{BC-AD} < d,$$

respectively. Recall that Euler's ϕ-function $\phi(b)$ is for $b \in \mathbb{N}$ defined as the number of positive integers $a \le b$ that are coprime to b. It satisfies the identity

$$\phi(b) = b\sum_{d|b} \frac{\mu(d)}{d}.$$

Taking this into account, we find

$$\Sigma_{\Delta}(n) = \frac{1}{2}\Delta^2 \sum_{1\le b\le n} \frac{\phi(b)}{b} \sum_{b\frac{D}{B} < d \le \min\{n, b\nabla\}} \frac{\phi(d)}{d}(bD - dB)^2, \tag{10}$$

$$\Sigma_{\diamondsuit}(n) = \frac{A}{B} \sum_{1\le b\le n} \frac{\phi(b)}{b} \sum_{b\nabla < d \le n} \frac{\phi(d)}{d}\left(bdB\Delta - b^2\left(D\Delta + \frac{1}{2}\frac{A}{B}\right)\right). \tag{11}$$

In order to evaluate these double sums, we have to compute the asymptotics for certain sums involving Euler's ϕ-function.

It is well-known that

$$\sum_{1\le b\le n} \phi(b) = \frac{3}{\pi^2}n^2 + O(n\log n).$$

By partial summation, it follows that

$$\sum_{N\le k\le M}\phi(k)k^{\ell}=\frac{6}{(2+\ell)\pi^2}\left(M^{2+\ell}-N^{2+\ell}\right)+O\left(M^{\ell+1}\log M\right)$$

for $\ell\ge -1$. Then, for $\ell\in\{0,\pm1\}$, we have

$$\begin{aligned}
&\sum_{1\le b\le n}\phi(b)b^{\ell}\sum_{xb<d\le n}\phi(d)d^{-\ell}\\
&\quad=\frac{6}{(2-\ell)\pi^2}\Bigg(n^{2-\ell}\sum_{1\le b\le\min\{n,n/x\}}\phi(b)b^{\ell}-x^{2-\ell}\sum_{1\le b\le\min\{n,n/x\}}\phi(b)b^{2}\Bigg)+\\
&\qquad+O\Big(n^{1-\ell}\log n\sum_{1\le b\le n}\phi(b)b^{\ell}\Big)\\
&\quad=\frac{n^4}{\pi^4}\Upsilon(x;\ell)+O\Big(n^3(\log n)^2\Big),
\end{aligned}$$

where

$$\Upsilon(x;\ell):=\frac{36}{4-\ell^2}\min\{1,x^{-2-\ell}\}-\frac{9}{2-\ell}\min\{x^{2-\ell},x^{-2-\ell}\}.$$

Similarly,

$$\begin{aligned}
&\sum_{1\le b\le n}\phi(b)b^{\ell}\sum_{\substack{1\le d\le n\\ xb<d\le yb}}\phi(d)d^{-\ell}\\
&\quad=\sum_{1\le b\le\min\{n,n/y\}}\phi(b)b^{\ell}\sum_{xb<d\le yb}\phi(d)d^{-\ell}+\\
&\qquad+\sum_{\min\{n,n/y\}<b\le\min\{n,n/x\}}\phi(b)b^{\ell}\sum_{xb<d\le n}\phi(d)d^{-\ell}\\
&\quad=\frac{6(y^{2-\ell}-x^{2-\ell})}{(2-\ell)\pi^2}\sum_{1\le b\le\min\{n,n/y\}}\phi(b)b^{2}+\\
&\qquad+\frac{6}{(2-\ell)\pi^2}\Bigg(n^{2-\ell}\sum_{\min\{n,n/y\}<b\le\min\{n,n/x\}}\phi(b)b^{\ell}+\\
&\qquad-x^{2-\ell}\sum_{\min\{n,n/y\}<b\le\min\{n,n/x\}}\phi(b)b^{2}\Bigg)+O\Big(n^{1-\ell}\log n\sum_{1\le b\le n}\phi(b)b^{\ell}\Big)\\
&\quad=\frac{n^4}{\pi^4}\Psi(x,y;\ell)+O\Big(n^3(\log n)^2\Big),
\end{aligned}$$

where

$$\begin{aligned}
\Psi(x,y;\ell)\;:=\;&\frac{9}{2-\ell}\Big(\min\{y^{2-\ell},y^{-2-\ell}\}-\min\{x^{2-\ell},x^{-2-\ell}\}\Big)+\\
&+\frac{36}{4-\ell^2}\Big(\min\{1,x^{-2-\ell}\}-\min\{1,y^{-2-\ell}\}\Big),
\end{aligned}$$

valid also for $\ell\in\{0,\pm1\}$. Using this in (10) and (11), it yields

$$\begin{aligned}
\Sigma_{\Delta}(n)&=\frac{n^4}{\pi^4}\frac{1}{2}\Delta^2\Big(D^2\Psi(1)-2BD\Psi(0)+B^2\Psi(-1)\Big)+O\Big(n^3(\log n)^2\Big),\\
\Sigma_{\diamondsuit}(n)&=\frac{n^4}{\pi^4}\frac{A}{B}\Big(B\Delta\Upsilon(0)-\Big(D\Delta+\frac{1}{2}\frac{A}{B}\Big)\Upsilon(1)\Big)+O\Big(n^3(\log n)^2\Big),
\end{aligned}$$

where $\Upsilon(\ell) = \Upsilon(\nabla;\ell)$ is given by (2) and $\Psi(\ell) = \Psi\Big(\frac{D}{B}, \nabla; \ell\Big)$ by (3). For the error term in (8) we note that in all cases

$$\mathsf{length}(\mathcal{R}(b,d)) \ll b\frac{A}{B} + d\frac{C}{D}$$

and this leads to the estimate $\mathsf{error}(n) \ll n^3$ in (9). Hence, we finally obtain

$$\begin{aligned}\Sigma(n) &= \frac{n^4}{\pi^4}\Big\{\frac{1}{2}\Delta^2\Big(D^2\Psi(1) - 2BD\Psi(0) + B^2\Psi(-1)\Big) + \\ &\qquad + \frac{A}{B}\Big(B\Delta\Upsilon(0) - \Big(D\Delta + \frac{1}{2}\frac{A}{B}\Big)\Upsilon(1)\Big)\Big\} + \\ &\quad + O\Big(n^3(\log n)^2\Big).\end{aligned}$$

Dividing this by the quantity in (6), we get the required proportion, and conclude the proof of the theorem.

3. Concluding remarks

The rather complicated expression for $\delta = \delta(\frac{A}{B}, \frac{C}{D})$ was checked by computer experiments. It is interesting to compare the values of δ for different data A, B, C, D. For instance,

$$\begin{aligned}\delta\left(\frac{1}{2}, \frac{1}{1}\right) &= \frac{11}{216} &&= 0.05092\ldots\ , \\ \delta\left(\frac{1}{3}, \frac{2}{3}\right) &= \frac{1}{72} &&= 0.01388\ldots\ , \\ \delta\left(\frac{1}{7}, \frac{1}{2}\right) &= \frac{1597}{6615} &&= 0.24142\ldots\ .\end{aligned}$$

In each of these three cases, computing the corresponding proportions $\Sigma(n)/(\Sigma(n;\frac{A}{B})\Sigma(n;\frac{C}{D}))$ for $n = 200$, we have obtained values that differ from the corresponding proportion δ by less than $7 \cdot 10^{-4}$.

It is also interesting to ask for a *global* proportion, i.e., when $0 < \frac{A}{B} < \frac{C}{D} \le 1$ are chosen randomly. However, with respect to this question our approach seems to be rather technical. We have computed the corresponding quantity

$$\lambda(n) := \frac{\sharp\Big\{\Big(\frac{a}{b}, \frac{A}{B}, \frac{c}{d}, \frac{C}{D}\Big) \in \mathcal{F}_n^4 \,:\, \frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D} \quad \text{and} \quad \frac{a+c}{b+d} > \frac{A+C}{B+D}\Big\}}{\sharp\Big\{\Big(\frac{a}{b}, \frac{A}{B}, \frac{c}{d}, \frac{C}{D}\Big) \in \mathcal{F}_n^4 \,:\, \frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D}\Big\}}$$

for several values of n and obtained the following values:

n	10	20	30
$\lambda(n)$	0.03755...	0.03750...	0.03847...

We conjecture that the limit of $\lambda(n)$ for $n \to \infty$ exists and is positive, probably around the value 0.04; however, we did not succeed in proving that and leave it as an open problem.

Acknowledgments

The authors are grateful to the anonymous referee for his or her valuable remarks and corrections to a first version of this article.

References

[1] M.R. Cohen, E. Nagel, *An Introduction to Logic and Scientific Method*, New York: Harcourt, Brace and Co 1934

[2] L.R. Ford, Fractions, *Amer. Math. Monthly* **45** (1938), 586-601

[3] G.H. Hardy, E.M. Wright, *An introduction to the theory of numbers*, Oxford University Press 1979, 5th ed.

[4] M.N. Huxley, *Area, lattice points, and exponential sums*, The Clarendon Press, Oxford University Press, New York 1996

[5] Y. Mittal, Homogeneity of subpopulations and Simpson's paradox, *J. Amer. Stat. Assoc.* **86**, No. 413 (1991), 167-172

[6] E.H. Simpson, The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society, Series B* **13** (1951), 238-241