INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

THE FUSSLER SAMPLING TECHNIQUE FOR POPULATIONS WITH A DISCRETE OR A CONTINUOUS DISTRIBUTION OF THICKNESSES

L. Egghe

LUC, Universitaire Campus, B-3610 Diepenbeek, Belgium (*)
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

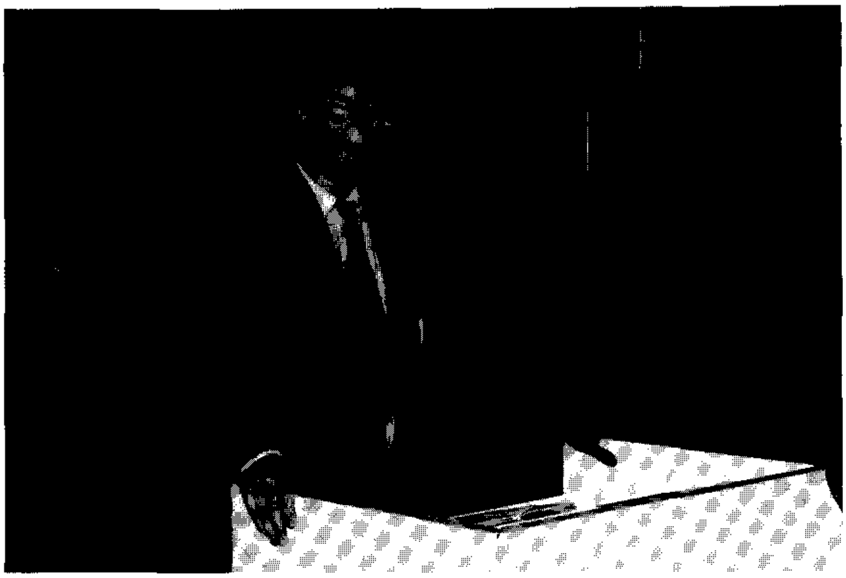
Abstract

In this paper we show that the Fussler sampling technique in book shelves is always better than systematic sampling by length. So far this result was only known to be true in the idealized situation of two categories of books : "thin" and "thick" books (Bookstein, Rousseau). In the present paper we allow any distribution of thicknesses of books on the shelf and furthermore we show that the same result is true for systems with a continuous distributions of thicknesses, which has applications in sampling by time.

1. INTRODUCTION

When sampling in book shelves in a large library, one can use several methods. The ideal sampling technique is random sampling (where the books are picked according to a table of random numbers), but this technique is extremely time consuming in this context. A fast, method is the sampling technique by length (which may be taken at random), but here a definite bias is introduced in favoring thick books : a book that has a thickness of twice the thickness of another book obviously has a double chance to be picked in a sampling technique by length, than has the second book.
Therefore, in an effort to combine fastness of the sampling technique together with randomness (as much as possible), Fussler introduced the following technique (see [5]) : execute a sampling by length but do not take the book that is choosen by this sampling act but take the k^{th}-book after it. Here k may be 1,2,3, ... but not high in order to have a fast technique (one may just take k=1 : the next book). Obviously, the Fussler sampling technique is as quick as the sampling technique by length.
The quality of the Fussler sampling technique has been investigated in [1] and [6]. Bookstein uses the model of sampling in a card file where one has the (idealized) situation of cards that can only have two possible thicknesses. Everything depends now on the way the "thin" and "thick" cards are clustered. If we denote by t a thin card and by T a thick card, one can measure this clustering by counting the number of groups of t's and T's in a file. For instance the clustering ttTTTTtTTtttttTTttTt has 5 groups of consecutive t's and 4 groups of consecutive T's, hence altogether R = 9 runs. The distribution of these runs can be shown (using cell occupancy theory, see e.g. [4], p. 42-43) to be approximately normal, as indicated in figure 1.
Bookstein in [1] studied only the left hand side of the above figure and showed in this case that, no matter how the t's and T's are clustered, one has less bias using the Fussler sampling technique than with sampling by length. Indeed, if P_1

(*) Permanent Address.



L. Egghe

denotes the chance to pick a thin card at random, P_2 denotes the chance to pick a thin card by using the Fussler sampling procedure (using any $k = 1,2,3,\ldots$ say $k = 1$), Bookstein shows that for clusters belonging to the left hand side of figure 1, we always have :

$$P_2 \leqslant P_3 \leqslant P_1 \qquad (1)$$

(and of course, the inequalities reverse for sampling thick cards) : so the Fussler technique is always closer to the random sampling technique than is the sampling technique by length.
The right hand side however has an equal chance to occur as the left hand side of figure 1. As was shown in [6], one even has here that $P_2 < P_3 < P_1$, but Rousseau shows that the following is true for all types of clustering of thin and thick cards :

$$|P_1 - P_3| \leqslant P_1 - P_2 \qquad (2)$$

showing that the Fussler technique is always better than sampling by length. Also, for the most common cases (the central part of fig. 1) we have that $P_1 \approx P_3$ (see fig. 2).

Furthermore, Rousseau uses the terminology of thin and thick books in a book shelve, where one calls a book thin if its thickness is smaller than or equal to a certain fixed number; otherwise it is called thick. Here then (2) is valid, denoting by P_i ($i = 1,2,3$), the chance to pick a thin book according to the random technique (P_1), sampling by length (P_2), or by using the Fussler technique (P_3). Note that, if we denote by $P'_i = 1 - P_i$ ($i = 1,2,3$), we have the chances to pick a thick book according to the three sampling techniques mentioned above. Here we have obviously :

$$|P'_1 - P'_3| \leqslant P'_2 - P'_1 \qquad (3)$$

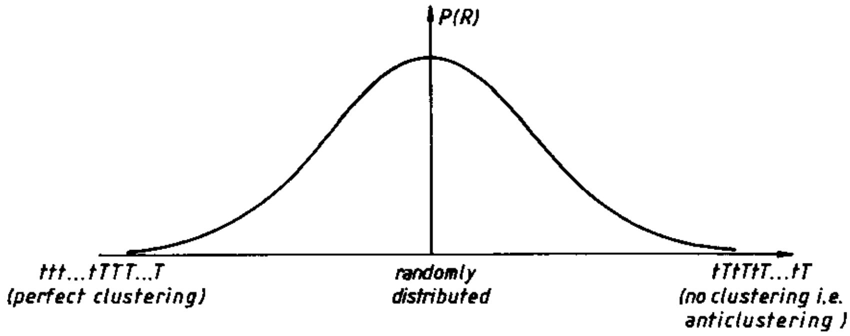

P(R)
ttt...tTTT...T
(perfect clustering)
randomly
distributed
tTtTtT...tT
(no clustering i.e.
anticlustering)

Fig. 1

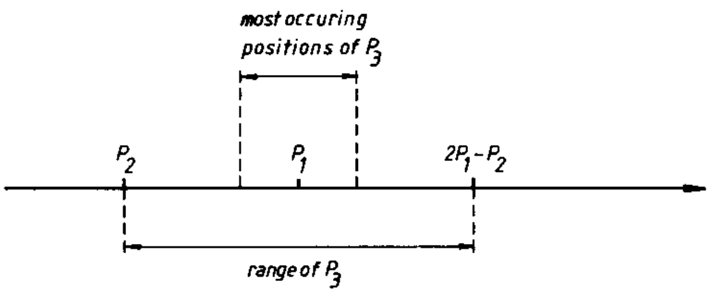

most occuring
positions of P_3
P_2
P_1
$2P_1 - P_2$
range of P_3

Fig. 2

All these results do not deal with the realistic situation of more than two types of thicknesses in a book shelve. In fact in reality we have a finite number of possible thicknesses of books, say (in increasing order) :

$$d_1 < d_2 < d_3 < \ldots < d_n . \qquad (4)$$

Denote by

$P_1(d_j)$ = the chance to pick a book with thickness d_j, using random sampling.

$P_2(d_j)$ = the chance to pick a book with thickness d_j, using sampling by length.

$P_3(d_j)$ = the chance to pick a book with thickness d_j, using the Fussler sampling technque (i.e. sampling by length as above and taking the k^{th} book after it, $k \in \{1,2,3,\ldots\}$ arbitrary but fixed).

In the next section, we will show that, for all j = 1, ..., n one has :

$$| P_1(d_j) - P_3(d_j)| \leq | P_1(d_j) - P_2(d_j) | \qquad (5)$$

showing that the Fussler sampling procedure is always better than sampling by length, no matter how the books of different thicknesses are clustered on the shelve. Note that, in view of inequalites (2) and (3), we need the absolute value signs in both sides of the inequality. Inequality (5) hence gives the final conclusion that the Fussler sampling technique is to be preferred if one wants to take a quick sample in book shelves.

In the last section we extend the above result to the case of a continuous distribution of "thicknesses". Besides the possible theoretical interest of this case, this result can be used in sampling by length of time, for instance in sampling how much books a library user checks out per library visit : the check-out time is a continuous random variable; hence, when taking samples in time, it is then better, based on our results in the last section, to take the next person waiting in line at a check-out desk, in order to avoid a bias towards more books checked out in one time. Another application might be in sampling computer-outputs.

2. FUSSLER SAMPLING IN BOOK SHELVES : THE CASE OF A DISCRETE DISTRIBUTION OF THICKNESSES

We use the notations of the previous section.

Theorem 1 : For every j = 1, ..., n, one has :

$$| P_1(d_j) - P_3(d_j)| \leq | P_1(d_j) - P_2(d_j)| \qquad (5)$$

Proof : Fix any j = 1, ..., n. In the set $\{d_j, d_{j+1}, \ldots, d_n\}$, books with thickness d_j are thin, in the sense of Rousseau [6] : here we consider - as Rousseau did - two types of books : the books with thickness d_j (the thin books) and books with thickness in $\{d_{j+1}, \ldots, d_n\}$ (the thick books). Hence, inequality (2) yields (now using conditional expectations, to be in the range $\{d_j, \ldots, d_n\}$)

$$| P_1(d_j | \{d_j, \ldots, d_n\}) - P_3(d_j | \{d_j, \ldots, d_n\})|$$
$$\leq P_1(d_j | \{d_j, \ldots, d_n\}) - P_2(d_j | \{d_j, \ldots, d_n\}) \qquad (6)$$

Using the definition of conditional expectations, we find :

$$\left|\frac{P_1(d_j)}{P_1(\{d_j, ..., d_n\})} - \frac{P_3(d_j)}{P_3(\{d_j, ..., d_n\})}\right| \leqslant \frac{P_1(d_j)}{P_1(\{d_j, ..., d_n\})} - \frac{P_2(d_j)}{P_2(\{d_j, ..., d_n\})} \quad (7)$$

or

$$\left|\frac{P_1(d_j)}{\sum_{\ell=j}^{n} P_1(d_\ell)} - \frac{P_3(d_j)}{\sum_{\ell=j}^{n} P_3(d_\ell)}\right| \leqslant \frac{P_1(d_j)}{\sum_{\ell=j}^{n} P_1(d_\ell)} - \frac{P_2(d_j)}{\sum_{\ell=j}^{n} P_2(d_\ell)} \quad (8)$$

Likewise, books with thickness d_j are thick in the range of books with thickness $\{d_1, ..., d_j\}$ (the books with thickness in the set $\{d_1, ..., d_{j-1}\}$ being considered as the thin books in the sense of Rousseau [6]). So, by using inequality (3) we have now :

$$|P_1(d_j | \{d_1, ..., d_j\}) - P_3(d_j | \{d_1, ..., d_j\})| \leqslant P_2(d_j | \{d_1, ..., d_j\}) - P_1(d_j | \{d_1, ..., d_j\}) \quad (9)$$

As above, we now find :

$$\left|\frac{P_1(d_j)}{\sum_{\ell=1}^{j} P_1(d_\ell)} - \frac{P_3(d_j)}{\sum_{\ell=1}^{j} P_3(d_\ell)}\right| \leqslant \frac{P_2(d_j)}{\sum_{\ell=1}^{j} P_2(d_\ell)} - \frac{P_1(d_j)}{\sum_{\ell=1}^{j} P_1(d_\ell)} \quad (10)$$

To simplify further calculations, we adopt some new notations. Put :

$$\alpha_i = \sum_{\ell=j}^{n} P_i(d_\ell) \qquad (i = 1,2,3) \quad (11)$$

and :

$$a_i = P_i(x_j) \qquad (i = 1,2,3) \quad (12)$$

(since j is fixed we do not mention the index j in a_i and α_i).

So, in this new notation formulas (8) and (10) become :

$$\left|\frac{a_1}{\alpha_1} - \frac{a_3}{\alpha_3}\right| \leqslant \frac{a_1}{\alpha_1} - \frac{a_2}{\alpha_2} \quad (13)$$

and

$$\left|\frac{a_1}{a_1 + 1 - \alpha_1} - \frac{a_3}{a_3 + 1 - \alpha_3}\right| \leqslant \frac{a_2}{a_2 + 1 - \alpha_2} - \frac{a_1}{a_1 + 1 - \alpha_1} \quad (14)$$

From these inequalities, it follows that :

$$|a_1\alpha_3 - a_3\alpha_1| \leqslant a_1\alpha_3 - a_2\frac{\alpha_1\alpha_3}{\alpha_2} \quad (15)$$

and :

$$|a_1(1-\alpha_3) - a_3(1-\alpha_1)| \leqslant a_2\frac{(a_1+1-\alpha_1)(a_3+1-\alpha_3)}{a_2 + 1 - \alpha_2} - a_1(a_3+1-\alpha_3) \quad (16)$$

Hence, using a triangular inequality

$$
\begin{aligned}
&| P_1(d_j) - P_3(d_j)| \\
&= | a_1 - a_3| \\
&\leq | a_1\alpha_3 - a_3\alpha_1 | + | a_1(1 - \alpha_3) - a_3(1 - \alpha_1)| \\
&\leq a_2 \left\{ \frac{(a_1 + 1 - \alpha_1)(a_3 + 1 - \alpha_3)}{a_2 + 1 - \alpha_2} - \frac{\alpha_1\alpha_3}{\alpha_2} \right\} - a_1(a_3 + 1 - 2\alpha_3) \\
&= P_2(d_j) \frac{(P_1(d_j) + 1 - \sum_{\ell=j}^{n} P_1(d_\ell))((P_3(d_j) + 1 - \sum_{\ell=j}^{n} P_3(d_\ell))}{P_2(d_j) + 1 - \sum_{\ell=j}^{n} P_2(d_\ell)} \\
&\quad - \frac{\sum_{\ell=j}^{n} P_1(d_\ell) \sum_{\ell=j}^{n} P_3(d_\ell)}{\sum_{\ell=j}^{n} P_2(d_\ell)} - P_1(d_j) \left[P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell)\right] \qquad (17)
\end{aligned}
$$

We now accept a second order approximation :

$$P_2(d_j) P_1(d_{j'}) \approx P_2(d_j) P_2(d_{j'})$$

for all $j, j' = 1, \ldots, n$ (since $P_2(d_j)$ is small). Now inequality (17) becomes :

$$
\begin{aligned}
&| P_1(d_j) - P_3(d_j)| \\
&\leq P_2(d_j) \left[P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell)\right] \\
&- P_1(d_j) \left[P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell)\right] . \qquad (18)
\end{aligned}
$$

Put :

$$\alpha = P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell) \qquad (19)$$

Then (18) reads :

$$| P_1(d_j) - P_3(d_j)| \leq \alpha \, (P_2(d_j) - P_1(d_j)) \qquad (20)$$

Now :

$$
\alpha \begin{cases} = 1 - P_3(d_j) - 2 \sum_{\ell=j+1}^{n} P_3(d_\ell) & \text{if } j < n \\ = 1 - P_3(d_j) & \text{if } j = n \end{cases}
$$
$$\leq 1 \text{ in all cases} . \qquad (21)$$

Furthermore, since

$$1 = \sum_{\ell=1}^{n} P_3(d_\ell) \geq \sum_{\ell=j}^{n} P_3(d_\ell)$$

we have that :

$$\alpha \geqslant 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell) \geqslant -1 \quad \text{in all cases .} \tag{22}$$

From (21) and (22) it now follows that

$$|\alpha| \leqslant 1 \tag{23}$$

in all cases. Inequalities (20) and (23) now imply

$$|P_1(d_j) - P_3(d_j)| \leqslant |P_1(d_j) - P_2(d_j)| \tag{5}$$

for every $j = 1, ..., n$. □

Remarks :

1. From formula (19) we see that, if d_j is small (the thinner books), $\alpha \approx -1$. For these books we have, using inequality (20), that

   $$|P_1(d_j) - P_3(d_j)| \leqslant P_1(d_j) - P_2(d_j) \tag{24}$$

   If d_j is large (the thicker books), we find $\alpha \approx 1$ and hence, using (20) :

   $$|P_1(d_j) - P_3(d_j)| \leqslant P_2(d_j) - P_1(d_j) \tag{25}$$

2. $|P_1(d_j) - P_2(d_j)|$ is large for d_j small or d_j large. Obviously, for average values of d_j, this quantity is small. But no matter how large $|P_1(d_j) - P_2(d_j)|$ is, inequality (5) is valid and it might be that, when the books of thickness d_j are randomly distributed amongst the other books (which is likely to be the case in book shelves), that $P_1(d_j) \approx P_3(d_j)$ even when $|P_1(d_j) - P_2(d_j)|$ is large. So the Fussler sampling technique has its strongest power in eliminating the largest bias (for d_j small or d_j large) encountered when sampling by length.

3. The sign of $P_1(d_j) - P_3(d_j)$ depends on the degree of clustering of the books with thickness d_j.
   In any case, inequality (5) shows that the Fussler sampling technique is better than sampling by length (but is as quick as the latter procedure).

4. The Fussler sampling technique is of course also applicable in card files in which one has cards of different thicknesses. This situation is equivalent with sampling in book shelves. Concerning this application, see also the remark 2 in the next section.

3. FUSSLER SAMPLING IN THE CASE OF A CONTINUOUS DISTRIBUTION FUNCTION

In this section we will extend the result of the previous section to the case of a continuous distribution function of "thicknesses". Since the applications of this are more in the area of sampling occurences of phenomena in time we will hence - forth speak of a continuous distribution of time.
In the introduction we already mentioned the application of sampling the number of books that are checked out (at the same time) in one library visit of a person. Therefore we will adopt this terminology, to fix the ideas (after this theory we will give some other possible applications).
Suppose the times to serve a library user, in checking out the books he/she wants to borrow varies between $t = 0$ and $t = t_m = t_{max}$. Service time is indeed a continuous random variable (often a negative exponential distribution).

In sampling the number of books library users check out in one time one might take a random sample in the populations of the users (situation 1). One might also taken a sample by time (here time is a random number) (situation 2); this method gives a bias towards the cases requiring a longer service time, hence towards the cases that more books are checked out by one person in one time. Situation 3 refers to situation 2 but we take the next borrower that is waiting in the line ($k = 1$ here; in general $k \in \mathbb{N}_0$): this is the Fussler technique, in essence.
Let $t_0, t_1 \in [0,t_m]$, $t_0 < t_1$ arbitrary. Denote then for $i = 1,2,3$, $P_i[t_0,t_1]$ = the probability to pick a borrower with a check-out time in the interval $[t_0,t_1]$, where the sampling method is as described in situation i above.

Theorem 2 : For every $t_0,t_1 \in [0,t_m]$, $t_0 < t_1$, one has :

$$| P_1[t_0,t_1] - P_3[t_0,t_1] | \leq | P_1[t_0,t_1] - P_2[t_0,t_1] | \qquad (26)$$

Proof : The proof follows the lines of the proof of theorem 1, remarking now that times in $[t_0,t_1]$ are short w.r.t. the time interval $[t_0,t_m]$ and that times in $[t_0,t_1]$ are long w.r.t. the time interval $[0,t_1]$. □

This shows that the Fussler sampling procedure is to be preferred above sampling by time, but is as quick as the latter procedure.
The probabilities $P_i[t_0,t_1]$ can be interpreted as probability measures on $\mathcal{B}[0,t_m]$, the Borel sets on the interval $[0,t_m]$. Denote by λ the Lebesgue-measure on $[0,t_m]$ (cf. [3]). We have the functional relations ($i = 1,2,3$) :

$$P_i : [t_0,t_1] \in \mathcal{B}[0,t_m] \to P_i[t_0,t_1]$$

Furthermore we have that $P_i \ll \lambda$ (i.e. P_i is absolutely continuous w.r.t. λ, which means that $\lambda(A) = 0$ implies $P_i(A) = 0$). Hence we can apply the theorem of Lebesgue-Radon-Nikodym (see [3], from p. 52 on). This theorem shows the existence of a function f_i ($i = 1,2,3$)

$$f_i : [0,t_m] \to \mathbb{R}$$

which is Lebesgue integrable, such that for every $A \in \mathcal{B}[0,t_m]$ and every $i = 1,2,3$

$$P_i(A) = \int_A f_i(t)\, dt \qquad (27)$$

In particular, for every $t_0, t_1 \in [0,t_m]$, $t_0 < t_1$ we also have

$$P_i[t_0,t_1] = \int_{t_0}^{t_1} f_i(t)\, dt \qquad (28)$$

The function f_i is called the density function of P_i w.r.t. λ.

We have the following result :

Theorem 3 :

$$| f_1(t) - f_3(t) | \leq | f_1(t) - f_2(t) |, \ \lambda\text{ - a.e.} \qquad (29)$$

Proof : From inequality (26) and formula (28) it follows that, for every $t_0,t_1 \in [0,t_m]$, $t_0 < t_1$:

$$\left| \int_{t_0}^{t_1} f_1(t)\,dt - \int_{t_0}^{t_1} f_3(t)\,dt \right| \leqslant \left| \int_{t_0}^{t_1} f_1(t)\,dt) - \int_{t_0}^{t_1} f_2(t)\,dt \right| \qquad (30)$$

t_1 can be written as $t_0 + h$ with $h > 0$. Furthermore, for every Lebesgue-integrable function f, one has :

$$\lim_{\substack{h \to 0 \\ >}} \frac{1}{h} \int_{t_0}^{t_0+h} f(t)\,dt = f(t_0), \ \lambda \text{ - a.e.} \qquad (31)$$

(see [3], p. 52). Since inequality (30) implies :

$$\left| \frac{1}{h} \int_{t_0}^{t_0+h} f_1(t)\,dt - \frac{1}{h} \int_{t_0}^{t_0+h} f_3(t)\,dt \right| \leqslant \left| \frac{1}{h} \int_{t_0}^{t_0+h} f_1(t)\,dt - \frac{1}{h} \int_{t_0}^{t_0+h} f_2(t)\,dt \right|$$

for every $t_0 \in [0,t_m]$ and $h > 0$ such that $t_0 + h \in [0,t_m]$, we have, by formula (31) :

$$| f_1(t) - f_3(t)| < | f_1(t) - f_2(t)| , \ \lambda \text{ - a.e.} \qquad (29)$$

Remarks

1. The results in this section do not only have applications in sampling check-outs. In the same way one may apply a quick Fussler procedure in sampling a computer output (of f.i. references obtained as a result of an online information retrieval search).

2. Another application can be found in returning to the situation described by Bookstein [1]. Indeed, as mentioned by Buckland, Hindle and Walker in [2], different tensions in different parts of a card drawer, furthermore dependent in time, can result in a continously changing "real" thickness (including the air between the cards).

The general conclusion is that all these uncontrollable physical aspects do not bother us : the Fussler sampling procedure is as quick as the sampling technique by length (or time) but gives less bias, and in fact reduces, in the most common cases to random sampling.

REFERENCES

[1] Bookstein, A., Sampling from card files. Libr. Q, 53, (1983) p. 307-312.
[2] Buckland, M.K., Hindle, A. and Walker, G.P.M., Methodological problems in assessing the overlap between bibliographic files and library holdings. Inf. Proc. and Manag. 11, (1975) p. 89-105.
[3] Burkill, J.C., The Lebesgue integral. Cambridge Tracts in Mathematics and Mathematical Physics, 40, (Cambridge University Press, 1971).
[4] Feller, W., An introduction to probability theory and its applications. Volume I, third edition, (John Wiley & Sons, Inc., 1957).
[5] Fussler, H., and Simon, J., Patterns in the use of books in large research libraries. (University of Chicago Press, 1961).
[6] Rousseau, R., The Fussler sampling technique. To appear in J. Amer. Soc. Inf. Sci.