Analysis of a dataset for Statistical Disclosure Control by random partition of a multi-index

Masaaki Sibuya (Keio U.) and Shido Sai (Okayama Shoka U.)

sibuyam@1986.jukuin.keio.ac.jp and shidosai@osu.ac.jp
Cherry Bud Workshop, 2008-03-28

Partition of multi-index

Counting data of many categories of small size

Number of categories is uncertain: (math "countably infinite")
Possible upper limit is determined by data.

Observation : size index $(s_1, s_2, \dots)$(frequency of frequencies, frequency spectrum)

Model : Random partition of a number.
Typical parameetric model is Ewens-Pitman sampling formula, EPSF.

Extension of random partition model

Sometimes the same type observations are repeated several times. The counts are classified into the same set of categories.

Random partition of a multi-index, or a vector of positive integer.

Model : multi-index extension of Ewens-Pitman samplinf formula, miEPSF.

Example. Analysis of Sai's data of US Cencus. (bi-partition)

An example of multi-index partition

$\nu = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$ partition of $|\nu| = 7 : 7, 6+1, 5+2, \cdots, 4+1+1+1, 3+2+1+1, 2+2+2+1, \dots$
Given partition $|\nu| = 3 + 2 + 1 + 1$,

4	3	1	0	0
3	0	1	1	1
7	3	2	1	1

$$\begin{bmatrix} 3 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 2 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix},$$

$$\begin{bmatrix} 2 & 0 & 1 & 1 \\ 1 & 2 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 2 & 1 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix}$$

size index of multi-index partition

1st \ 2nd	0	1	2	3	$s_{\cdot j}$	$js_{\cdot j}$
0	-	2	0	0	2	0
1	0	1	0	0	1	1
2	0	0	0	0	0	0
3	1	0	0	0	1	3
$s_{i\cdot}$	1	3	0	0	4	4
$is_{i\cdot}$	0	3	0	0	3	7

$$\begin{bmatrix} - & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} - & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} - & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\begin{bmatrix} - & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} - & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} - & 0 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} - & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Remark. Size index may concentrate in the 1st row or 1st column (partition #2, #8). This happens, if number of columns is large in the contingency table. This does not happen if the rowwise sums ¿1.

If the partition of $7 = 7, 6+1, 5+2, \cdots$ ($\pi(7) = 15$ ways) is given, the possibe number of multi-index $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$ is as follows.

prt.	#m.i.	prt.sum	prt.	#m.i.	prt.sum	prt.	#m.i.	prt.sum	prt.	#m.i.	prt.sum
7	1	1	511	3		4111	4		22111	5	9
61	2		421	6		3211	8		211111	3	3
52	3		331	4		2221	4	16	1111111	1	1
43	4	9	322	5	18	31111	4				
									total sum	57	57

Any probability measure on these 57 mi-partition is a random partition of $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$

Random partition of multi-index

In miEPSF, discussed later, given marginal partitions $\nu = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$, $l = [3, 2, 1, 1]$,

$$n = 7 = 3 + 2 + 1 + 1, \quad s = (2, 1, 1), \quad \pi_S(s, n) = \frac{7!}{2!(1!)^2 1!(2!)1!(3!)} = 210$$

partition#	1	2	3	4	5	6	7	8	sum
coef.	12	12	18	72	18	36	36	6	210
probability	$\frac{2}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{6}{35}$	$\frac{3}{35}$	$\frac{6}{35}$	$\frac{6}{35}$	$\frac{1}{35}$	1

Number 4 and Number 8 partitions:

$$\begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}, \ \frac{4!3!}{1!(2!)1!1!1!1!} = 72; \quad \begin{bmatrix} 0 & 2 & 1 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix}, \ \frac{4!3!}{1!(3!)1!(2!)2!(1!)1!1!} = 6$$

Expected size index $\times 210$ of bi-partition is

$s_1 \backslash s_2$	0	1	2	3	$s_{\cdot j}$	$js_{\cdot j}$
0	-	180	30	6	216	0
1	240	120	72	0	432	432
2	60	108	0	0	168	336
3	24	0	0	0	24	72
$s_{i\cdot}$	324	408	102	6	840	840
$is_{i\cdot}$	0	408	204	18	630	1470

expected size $\times 210$ of $s_1 + s_2 = 1, 2, 3$: 420, 210, 210.

US Census data

US Decennial Census 1990, 2000, Public Use Microdata Sample (PUMS) 1%,
Census of Population and Housing, Washington State,
(1% , age$\geq$20, classified by 12 “key variables”) See the attached Tables 1 and 2.

cell	C_1	C_2		C_ν		sum
1990	x_{11}	x_{12}	$\cdots$	$x_{1\nu}$	$\cdots$	n_1
2000	x_{21}	x_{22}	$\cdots$	$x_{2\nu}$	$\cdots$	n_2

$$s_{ij} = \sum_{\nu=1}^{\infty} I[x_{1\nu} = i \ \& \ x_{2\nu} = j], \ s_{i\cdot} = \sum_{j=1}^{n_2} s_{ij}, \ s_{\cdot j} = \sum_{i=1}^{n_1} s_{ij}, \ i, j \in \mathbb{Z}_{\geq 0}$$

$$s_k = \sum_{i+j=k} s_{ij}, \ s_{\cdot\cdot} = \sum_{i=0}^{n_1} s_{i\cdot} = \sum_{j=0}^{n_2} s_{\cdot j} = \sum_{k=1}^{n_1+n_2} s_k,$$

$$\sum_{i=1}^{n_1} is_{i\cdot} =: n_1, \ \sum_{j=1}^{n_2} js_{\cdot j} =: n_2, \ \sum_{i=0}^{n_1}\sum_{j=0}^{n_2}(i+j)s_{ij} = n_1 + n_2 =: n.$$

s_k is size index of $(x_{1\nu} + x_{2\nu})_{\nu=1}^{\infty}$
$s_{i\cdot}$, $i > 0$, is size index of $(x_{1\nu})_{\nu=1}^{\infty}$
$s_{\cdot j}$, $j > 0$, is size index of $(x_{2\nu})_{\nu=1}^{\infty}$

s_{00} is disregarded since random 0 and true 0 cannot be distinguished. $s_{0\cdot}$ and $s_{\cdot 0}$ are new type of statistics. Note that $s_0 = s_{00}$. $s_{\cdot\cdot}$is the sum of $x_{1\nu} + x_{2\nu} > 0$: total number of categories. $s - s_{0\cdot} = \sum_{i=1}^{n_1} s_{i\cdot}$, $s - s_{\cdot 0} = \sum_{j=1}^{n_2} s_{\cdot j}$are number of categories observed in 1990 and 2000, respectively.

The previous section shows the case where $(n_1, n_2) = (4, 3)$. The conditional distribution of the contingency table, or bi-partiton size index, is independent of the EPSF parameter (θ, α).

Analysis of census dataset A

1990: $s_{i\cdot}$ $(31919, 282, 57, 20, 14, 9, \ldots)$
2000: $s_{\cdot j}$ $(38314, 610, 103, 29, 21, 20, \ldots)$
combined: s_k $(69534, 1022, 185, 56, 34, 27, \ldots)$

summary statistics (marginal sums)

0	-	2000 only
-	containing both	1990 sum
1990 only	2000 sum	total number

cell numbers

1990\ 2000	0 +	sum	
0	0	-	$s_{0\cdot} = 38664$
+	-	$s_{++} = 525$	$s_{+\cdot} = 32387$
sum	$s_{\cdot 0} = 31862$	$s_{\cdot +} = 39189$	$s_{\cdot\cdot} = 71051$

$$s_{0\cdot} + s_{+\cdot} = s_{\cdot 0} + s_{\cdot +} = s_{\cdot\cdot}$$

$$s_{0\cdot} + s_{\cdot 0} + s_{++} = s_{+\cdot} + s_{\cdot +} - s_{++} = s_{\cdot\cdot}$$

individual numbers

1990\ 2000	0	+	sum
0	0	-	$n_{0\cdot} = 40311$
+	-	$n_{++} = 3541$	$n_1 = 34542$
sum	$n_{0\cdot} = 32699$	$n_2 = 41959$	$n. = 76501$

Ewens - Pitman sampling formula (EPSF) or Pitman's two-parameter random partition

$$\begin{array}{rrr} 1990 & \hat{\theta} = \ \ 169.3, & \hat{\alpha} = 0.9822, \\ 2000 & \hat{\theta} = 1499.6, & \hat{\alpha} = 0.9747, \\ \text{combined} & \hat{\theta} = 1500.0, & \hat{\alpha} = 0.9769. \end{array}$$

(For α close to 1, the value of θ does not effect much the likelihood. The fit is not good at tail in all three cases. See attached figures.)

Extension of EPSF to multi-index partition (miEPSF):
size index of combined partition of number is sufficient
random subdivision of numbers to multi-index is parameter-free

EPSF and Pólya urn model

Balls $B_1, B_2, \ldots$, are randomly and sequentially put into urns $U_1, U_2, \ldots$
Ball B_1 is put into U_1 with probability 1. If $B_1, \ldots, B_n$ are in $U_1, \ldots, U_k$, in such a way that $c_j > 0$ balls are in U_j, $j = 1, \ldots, k$, $\sum_{j=1}^k c_j = n$, ball B_{n+1} is put into

- a new urn U_{k+1} with probability $\frac{\theta+k\alpha}{\theta+n}$,
- an old urn U_j with probability $\frac{c_j-\alpha}{\theta+n}$, $1 \le j \le k$, $\quad 0 \le \alpha < 1$, $-\alpha < \theta < \infty$.

At the n-th stage when ball B_n is put, let S_j denote the number of urns occupied by j balls. $S = (S_1, \cdots, S_n)$ is the size index of a random partition of $n : \sum_{j=1}^n jS_j = n$, following the probability distribution, which is called Ewens-Pitman sampling formula.

$$
\begin{aligned}
P\{S = (s_1, \cdots, s_n)\} &= \frac{(\theta| - \alpha)_k n!}{(\theta| - 1)_n} \prod_{j=1}^n \frac{1}{s_j!} \left(\frac{(1-\alpha| - 1)_{j-1}}{j!} \right)^{s_j} \\
&= \frac{(\theta| - \alpha)_k}{(\theta| - 1)_n} \pi_S(s, n) \prod_{j=1}^n ((1-\alpha| - 1)_{j-1})^{s_j}, \qquad (1) \\
\pi_S(s, n) &= \frac{n!}{\prod_{j=1}^n s_j!(j!)^{s_j}}, \quad s \in \mathcal{P}_{nk} : k = \sum_{j=1}^n s_j, \ n = \sum_{j=1}^n j s_j.
\end{aligned}
$$

where $(t|a)_n = t(t-a)\cdots(t-(n-1)a)$, and $\mathcal{P}_{nk}$ is the set of all partitions of n to k terms. The range of the parameter $(0, \alpha)$ is

$$
\begin{aligned}
&0 \le \alpha < 1 \quad \text{and} \quad -\alpha < \theta < \infty, \\
\text{or} \quad &\alpha < 0 \quad \text{and} \quad \theta = -m\alpha, \quad m = 1, 2, \cdots
\end{aligned}
$$

Denote the probability distribution of S be denoted by EPSF(n; θ, α), and call its sequence $n = 1, 2, \cdots$ Pólya urn process written as $(\mathcal{S}_n)_{n=1}^\infty$.

A genesis of miEPSF

Table 1: genesis of miEPSF

Z_i	K	$T_K = \sum_{i=1}^{K} Z_i$	$(Z_1, Z_2, \dots)\vert T_K = n$
compounded dist.	compounding dist.	compound dist.	miEPSF
	ETNgBn	ETNgMn	$-\alpha < \theta < 0$ & $0 \leq \alpha < 1$
ETNgMn	TNgBn	TNgMn	$0 < \theta$ & $0 \leq \alpha < 1$
	LgSer	MvLgSer	$0 = \theta$ & $0 \leq \alpha < 1$
	TBn	TNgMn	$\theta = -m\alpha$ & $\alpha < 0$

Multi-index extension of EPSF

$$\begin{aligned}
P\{S = (s_{\underline{\iota}})\} &= \frac{(\theta| -\alpha)_k \, \nu!}{(\theta| -1)_{|\nu|}} \prod_{\iota} \frac{1}{s_\iota!} \left(\frac{(1-\alpha| -1)_{|\iota|-1}}{\iota!} \right)^{s_\iota} \\
&= \frac{(\theta| -\alpha)_k}{(\theta| -1)_n} \pi_S((s_{\underline{\iota}}), \nu) \prod_{\iota} ((1-\alpha| -1)_{|\iota|-1})^{s_\iota},
\end{aligned} \tag{2}$$

$$\pi_S(s_{\underline{\iota}}, \nu) = \frac{\nu!}{\prod_\iota s_\iota! (\iota!)^{s_\iota}}, \quad (s_{\underline{\iota}}) \in \mathcal{P}_{\nu k} : \sum_\iota s_\iota = k, \ \sum_\iota s_\iota \iota = \nu, \quad \nu! = \prod_{i=1}^{m} \nu_i! \, .$$

Theorem. Consider the joint distribution of a Pólya urn process $(\mathcal{S}_n)_{n=1}^{\infty}$ at $n = \nu_1, \ \nu_1 + \nu_2, \dots, \nu_1 + \cdots + \nu_d$.
The joint distribution of

$$T_{\nu_i} := \mathcal{S}_{\nu_1 + \cdots + \nu_i} - \mathcal{S}_{\nu_1 + \cdots + \nu_{i-1}}, \quad i = 1, \cdots, d, \quad \mathcal{S}_0 = 0,$$

is a miEPSF of the multi-index $\nu = \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_d \end{bmatrix}$.

Random subdivision of size index to obtain random partition of multi-index

random partition $s_{\underline{\iota}} \in \mathcal{P}_{\nu,k}$ of multi-index ν given "marginal" random size index $(s_1, s_2, \dots) \in \mathcal{P}_{n,k}$:

$$\frac{\nu!}{\prod_\iota s_\iota!(\iota!)^{s_\iota}} \Big/ \frac{n!}{\prod_{j=1}^n s_j!(j!)^{s_j}} = \prod_{j=1}^n \frac{s_j!}{\prod_{|\iota|=j} s_\iota!} \frac{(j!)^{s_j}}{\prod_{|\iota|=j}(\iota!)^{s_\iota}} \Big/ \frac{n!}{\nu!}$$
$$= \prod_{j=1}^n \left(\binom{s_j}{s_\iota} \prod_{|\iota|=j} \binom{j}{\iota}^{s_\iota} \right) \Big/ \binom{n}{\nu}, \quad n = |\nu|, \tag{3}$$
$$\binom{n}{\nu} = \binom{n}{\nu_1, \dots, \nu_m}.$$

If $m = 2$ and $s_i = 0,\ i \neq j$, that is $n = js_j$, let the size index of $\iota = (\ell, j-\ell)$ be denoted by z_ℓ instead of s_ι. Since

$$\sum_{\ell=0}^j \ell z_\ell = \nu_1 \text{ and } \sum_{\ell=0}^j (j-\ell) z_\ell = \nu_2, \quad \text{or equivalently,} \quad \sum_{\ell=0}^j z_\ell = s_j,$$

there are $j-1$ free variables, say $(z_1, \dots, z_{j-1})$. Because of $n = |\nu| = \nu_1 + \nu_2 = js_j$, ν satisfies one of the conditions $(\nu_1, \nu_2) \equiv (\ell, j-\ell) \pmod j$, $j = 0, 1, \dots, j-1$. Correspondingly, $(z_1, \dots, z_{j-1})$ may take j^{j-2} points of the hyper-cubic lattice $\{0, \dots, j-1\}^{j-1} \pmod j$. The joint pmf on these points is

$$\frac{s_j!}{\prod_{\ell=0}^j z_\ell!} \prod_{\ell=1}^{j-1} \binom{j}{\ell}^{z_\ell} \Big/ \binom{n}{\nu_1}, \qquad z_j = (\nu_1 - \sum_{\ell=1}^{j-1} \ell z_\ell)/j,\ z_0 = (\nu_2 - \sum_{\ell=1}^{j-1} (j-\ell) z_\ell)/j. \tag{4}$$

Its factorial moments are given by

$$E(\prod_\ell Z_\ell^{\underline{r_\ell}}) = s_j^{\underline{r}} \prod_\ell \binom{j}{\ell}^{r_\ell} \frac{\binom{n-jr}{\nu_1 - R}}{\binom{n}{\nu_1}} \quad r = \sum_\ell r_\ell,\ R = \sum_\ell \ell r_\ell.$$

For example,

$$E(Z_\ell) = s_j \binom{j}{\ell} \frac{\nu_1^{\underline{\ell}} \nu_2^{\underline{j-\ell}}}{n^{\underline{j}}}, \quad \text{and} \quad E(Z_\ell^{\underline{2}}) = s_j^{\underline{2}} \binom{j}{\ell}^2 \frac{\nu_1^{\underline{2\ell}} \nu_2^{\underline{2j-2\ell}}}{n^{\underline{2j}}}$$

In the simplest case of $j = 2$, returning to the old notation,

$$m = 2,\ n = s_2,\ \nu_1 + \nu_2 = 2s_2,\ \iota = (2,0), (1,1), (0,2),$$
$$\nu_1 = 2s_{20} + s_{11}, \quad \nu_2 = s_{11} + 2s_{02},$$
$$\frac{s_2! 2^{s_{11}}}{s_{20}! s_{11}! s_{02}!} \Big/ \binom{2s_2}{\nu_1} = \frac{s_2! \nu_1! \nu_2! 2^{s_{11}}}{(2s_2)! s_{11}! ((\nu_1 - s_{11})/2)! ((\nu_2 - s_{11})/2)!}, \tag{5}$$
$$\nu_1 \bmod 2 \le s_{11} \le \min(\nu_1, \nu_2).$$

The first and the second factorial moments of s_{11} are

$$\frac{\nu_1\nu_2}{n-1} \quad \text{and} \quad \frac{\nu_1(\nu_1-1)\nu_2(\nu_2-1)}{(n-1)!!}. \tag{6}$$

In general, without restriction $\nu_1 + \nu_2 = 2s_2$, (ν_1, ν_2) should be replaced by the random variable (m_1, m_2) where X follows the hypergeometric distribution with the marginals $(\nu_1, \nu_2; 2s_2, n-2s_2)$, and the moments (6) given marginals (ν_1, ν_2) and $(s_1, s_2, \dots)$ are

$$\frac{4s_2^2\nu_1\nu_2}{n^{\underline{2}}} \quad \text{and} \quad \frac{4s_2^{\underline{2}}\nu_1^{\underline{2}}\nu_2^{\underline{2}}}{n^{\underline{4}}}. \tag{7}$$

Sampling algorithm

Given size index $(s_1, \dots, s_n)$, $s_j \geq 0$, $\sum_{j=1}^n js_j = n$ and multi-index $\nu \in \mathbb{Z}^m_{>0}$, $|\nu| = n$
To generate a random partition of ν following (4);

1. From $(\nu_1, \dots, \nu_d)$ balls of d colors, take s_1 at random, and from the remainder take $2s_2$, and $3s_3$ and so on. The result is a two way contingency table with fixed marginals $(\nu_1, \dots, \nu_d)$ and $(js_j, 1 \leq j \leq n)$. The below left table.
   All possible contingency tables may appear. If there is no j such that $s_j > 1$, then the partition to $(j_1, \dots, j_n)$, $j_1 < \cdots < j_k$, $j_1 + \cdots + j_k = |\nu|$ is as usual contingency table.

2. Consider the j-th column with the marginal js_j. Suppose the number of balls of d colors be $m_1, \dots, m_d$, $m_1 + \cdots + m_d = js_j$. Forget for a while the colors of balls and mix the balls. Now subdivide the column to s_j columns with j balls. s_j columns are not distinguished. From $0, 1, \dots, js_j - 1$, take at random m, numbers, allocate these to the first color, and if the chosen number is x, put the ball to the column x modulo j. The below right table shows multi-index partition and its size index. This is *not* a contingency table.

1 st	ν_1	m_{11}	$\dots$	m_{1j}	$\dots$	m_{1n}
2 nd	ν_2	m_{21}	$\dots$	m_{2j}	$\dots$	m_{2n}
cmbd	n	s_1	$\dots$	js_j	$\dots$	ns_n

m_{1j}	0	1	$\dots$	j
m_{2j}	j	$j-1$	$\dots$	0
$(j\times)s_j$	s_{0j}	s_{1j}	$\dots$	s_{j0}

Analysis of census dataset A. continued

size index of combined partition of number is sufficient
random subdivision of numbers to multi-index is parameter free

Simulation given; (1) combined size index $(69534, 1022, 185, \cdots)$, and (2) number of individuals in 1990 and 2000 survey, $34552 + 41959 = 76501$.

simulation results (100 repetitions)

cell numbers

0	-	3847686
-	94323	3257414
3163091	3942009	**7105100**

individual numbers

0	-	3887177
-	574460	3454200
3188463	4195900	**7650100**

(marginal size indices of 1990 and 2000 are very similar)

cells containing surveyed of both year(actual/simulation)

cells numbers 943.23 vs 525 (0.56)

individuals 5744.60 vs 3541 (0.62)

Discussion

1. The filled cells change largely: Among 71,051 cells filled by the survey in both years, 31,862 0f 1990 disappeared and 38,664 appeared in 2000. Moreover, most of disappeared and appeared are cells of isotone:

	singleton	others	sum
1990	31551	311	31862
2000	37983	681	38664

   Hence, the change of numbers of individuals is similar.

2. Contrarily, the number of cells including observations of both year is very small: 525 cells (3541 persons). Almost all of them have small size.

size	2	272 cells	272 persons
	3	73 cells	146 persons

3. On the other hand, there are cells of large size, both in 1990 and 2000.
   Because of, my guess, a sort of cohort effect. Development of a new industry and a new town attracts working people. If 10 years pass without big immigration of emigration, that generation moves in mass to another cell.

Analysis of census dataset B

In the dataset A, it was found that the larger cells close to margins (containing mainly observed in either 1990 or 2000) are those of unemployed. Hence the second dataset B consists of employed only. The same computation and simulation are repeated for the new dataset, and the results are as follows.

summary statistics (marginal sums)

cell numbers

1990\ 2000	0 +	sum	
0	0	-	$s_{0\cdot} = 23204$
+	-	$s_{++} = 2390$	$s_{+\cdot} = 21359$
sum	$s_{\cdot 0} = 18969$	$s_{\cdot +} = 25594$	$s_{\cdot\cdot} = 44563$

$$s_{0\cdot} + s_{+\cdot} = s_{\cdot 0} + s_{\cdot +} = s_{\cdot\cdot}$$

$$s_{0\cdot} + s_{\cdot 0} + s_{++} = s_{+\cdot} + s_{\cdot +} - s_{++} = s_{\cdot\cdot}$$

individual numbers

1990\ 2000	0	+	sum
0	0	-	$n_{0\cdot} = 24899$
+	-	$n_{++} = 10192$	$n_1 = 24846$
sum	$n_{0\cdot} = 19989$	$n_2 = 30234$	$n_{\cdot} = 55080$

$$\begin{array}{lrl} 1990 & \hat{\theta} = & 4452.5, \quad \hat{\alpha} = 0.8746, \\ 2000 & \hat{\theta} = & 5292.0, \quad \hat{\alpha} = 0.8634, \\ \text{combined} & \hat{\theta} = & 5771.5, \quad \hat{\alpha} = 0.8657. \end{array}$$

simulation results (100 repetitions)

cell numbers

0	-	1872300
-	285648	258400
2298352	2157948	**4456300**

individual numbers

0	-	2418935
-	1143690	2484600
1945375	3023400	**5508000**

cells containing surveyed of both year(actual/simulation)

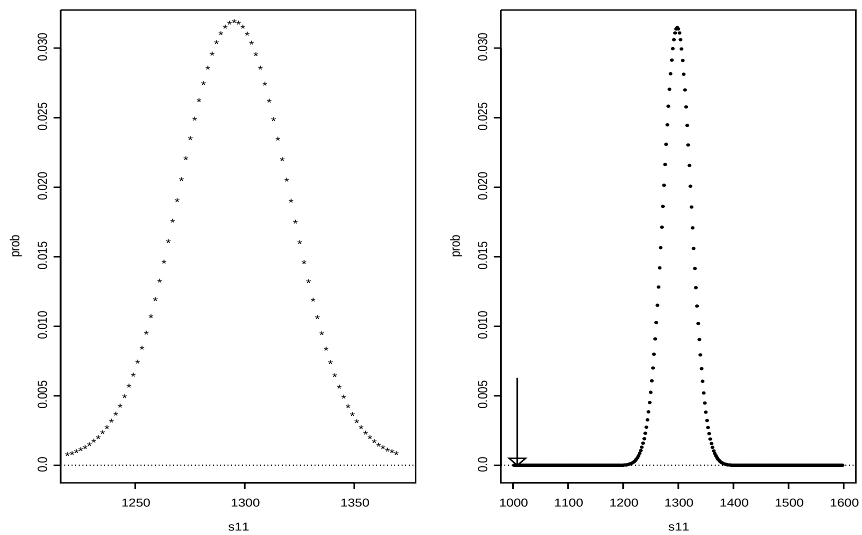
cells numbers 2856.48 vs 2390 (0.84)

individuals 11436.90 vs 10192 (0.89)

The fit of EPSF is satisfactory (attached figures), and the central part of s_{ij} is not sparse as the dataset A. However, computing the distribution and moments of s_{11}, (6) and (7), the observed $s_{11} = 1008$ is too small. The value is smaller than the expected value by 11.44×SD in (6) and 11.36×SD in (7).

equation	mean	SD	dev. of obs. ($\times$ SD)
(6)	1297.96	25.35	11.44
(7)	1298.48	25.58	11.36

The pmf of s_{11} (5) is plotted with the observed value in the figure below. The behavior of two theoretical distributions are very close as shown also in the above table. This is due to the fact that the observed ratio $(n_1, n_2) = 2358 : 2886$ is close to the marginal ratio $(\nu_1, \nu_2) = 24846 : 30234$.



Future work for Statistical Disclosure Control

- Why s_{++} is small. (my guess) A sort of cohort effect.
- s_{11} is small. Good news for SDC.
- Changing key variables, s_{++} will change. Our statistics will help to compare them.
- More flexible model to gain insight into dependence structure.

References

[1] Charalambides, C.A. (2005) *Combinatoric Methods in Discrete Distributions*, Wiley-Interscience, Hoboken, NJ.

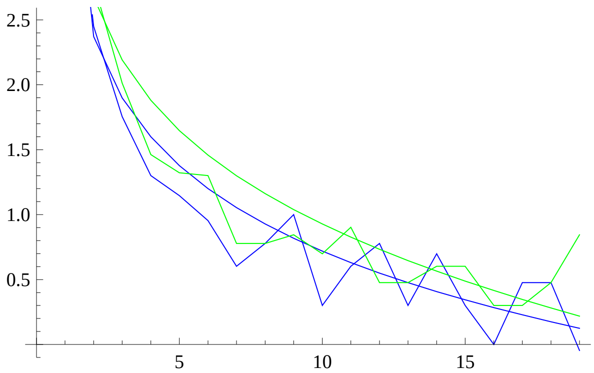
[2] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, Wiley.

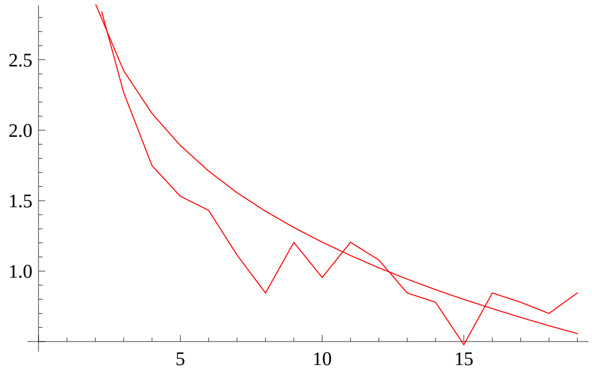
[3] Pitman, J. (2006) *Combinatorial Stochastic Processes*, Lecture Notes in Mathematics, **1875**, Springer, New York, NY.

```
CellPrint[{Cell["Dataset A", "Section"], Cell["EPSF fit: log size index", "Text"],
  Cell["EPSF fit: cumulative individuals", "Text"],
  Cell["Dataset B", "Section"], Cell["EPSF fit: log size index", "Text"],
  Cell["EPSF fit: cumulative individuals", "Text"]}]
```

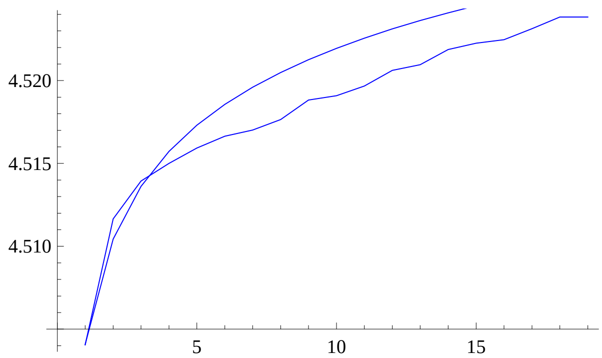
Dataset A

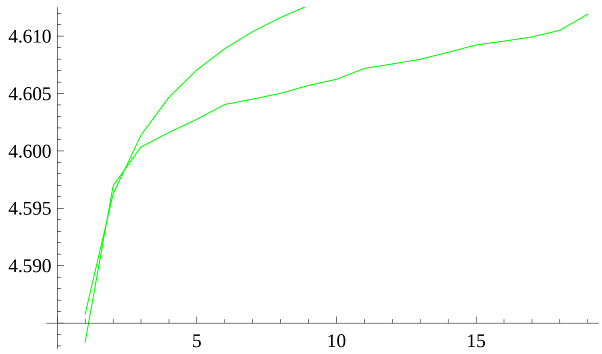
EPSF fit: log size index

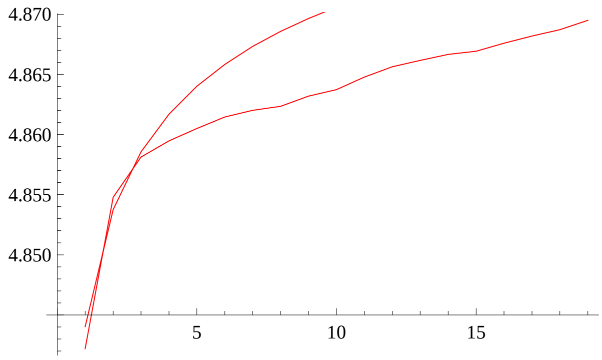




EPSF fit: cumulative individuals

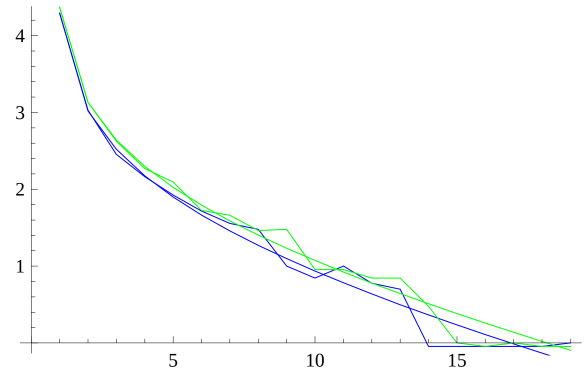


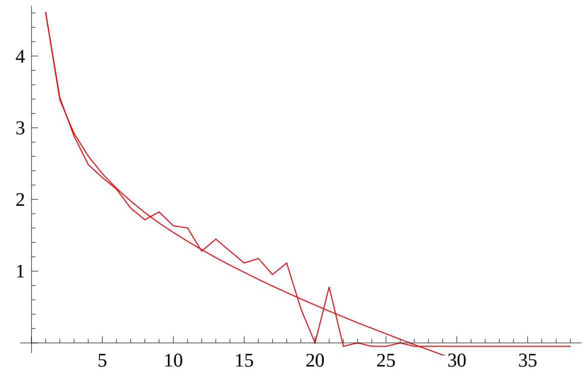




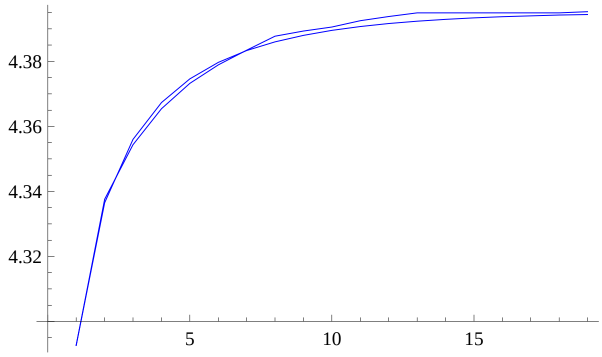
Dataset B

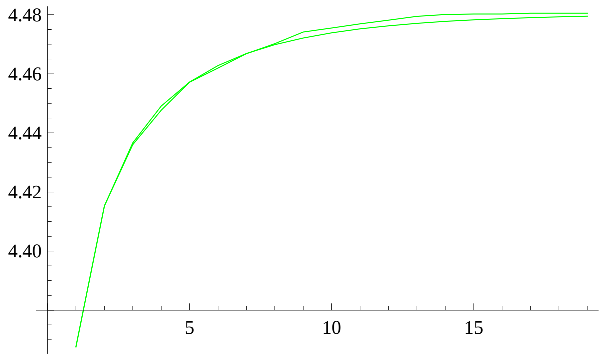
EPSF fit: log size index

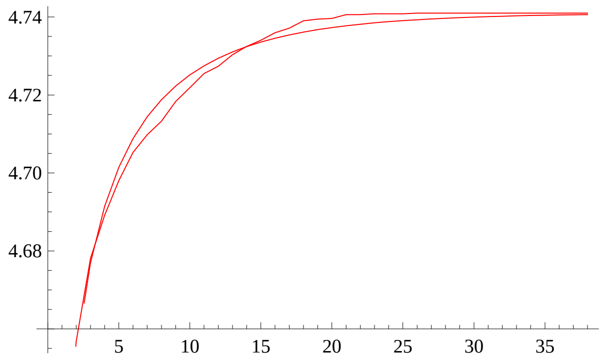




EPSF fit: log cumulative individuals


4.38
4.36
4.34
4.32
5
10
15


4.48
4.46
4.44
4.42
4.40
5
10
15


4.74
4.72
4.70
4.68
5
10
15
20
25
30
35

母集団多重寸法指標

1990年＼2000年	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	計
0		21971	939	202	53	23	6	4	1	5											23204
1	18154	1008	254	85	58	33	7	9	5	1	2	1									19617
2	675	202	78	48	18	20	8	8	3	4	2	1	2	1							1070
3	104	72	32	34	7	7	12	5	5	2	1	2		2							285
4	19	31	28	21	12	8	4	7	3	4	1	2	1	2	2			1			146
5	10	16	7	10	17	9	4	1	3	3		1	2	1							84
6	4	8	4	5	8	9	2	5	2	2		1	1	1							52
7	2	1	6	8	3	1	4	2	1	4	1	1	1		1						36
8			4	4	4	6	3	2	3	1	2					1					30
9	1	1		2	2	1	1			2											10
10		2	1	2		2															7
11				3	2	3		1		1											10
12					1	1	2	1		1											6
13					1	1			3												5
14																					0
15																					0
16																					0
17																					0
18																					0
19								1													1
計	18969	23312	1353	424	186	124	53	46	29	30	9	9	7	7	3	1	0	1	0	0	44563