

How to optimally coordinate two samples

Alina Matei



Université de Neuchâtel

Aarau, 19 November 2004

General Frame

- finite population $U = \{1, \dots, k, \dots, N\}$;
- two samples s_1, s_2 drawn in two different time occasions, with the probabilities $p_1(s_1), p_2(s_2), p(s_1, s_2)$;
- sample coordination (positive or negative coordination);
- inclusion probabilities $\pi_k^1 = Pr(k \in s_1), \pi_k^2 = Pr(k \in s_2)$ for all $k \in U$;
- $\pi_k^{1,2} = Pr(k \in s_1, k \in s_2)$ = joint inclusion probability of unit k in the first and second sample occasion.
- \mathcal{S}_1 and \mathcal{S}_2 the sample supports in the first and second occasions, respectively and $\#\mathcal{S}_1 = m$ and $\#\mathcal{S}_2 = q$.

Sample coordination

We have two kinds of sample coordination:

- positive coordination (when the goal is to maximize the number of common units of two or more samples);
- negative coordination (when the goal is to minimize the number of common units of two or more samples).

We focus on the positive coordination.

Example panel survey generally focus on the difference of the same characteristic (the difference in unemployment rate measured in time on 2 or more occasions).

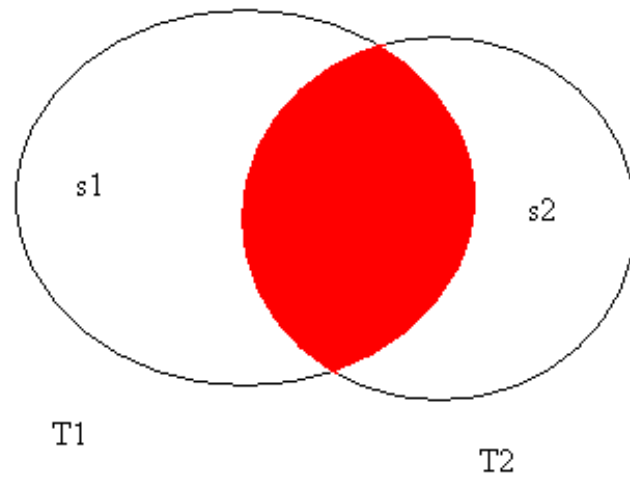


Figure 1: Positive coordination

Remarks

- The overlap between two samples is defined as the number of common units in both samples.
- Each unit $k \in U$ can be selected in both samples with probability at most

$$\min(\pi_k^1, \pi_k^2).$$

- An upper bound of the expected overlap is

$$\sum_{k \in U} \min(\pi_k^1, \pi_k^2).$$

- We call this bound the **absolute upper bound**. It is reached when

$$\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2), \text{ for all } k \in U.$$

- When 2 samples are drawn independently (without coordination): $\pi_k^1 \pi_k^2 = \pi_k^{1,2}$ for all $k \in U$.
- In positive coordination: $\pi_k^1 \pi_k^2 \leq \pi_k^{1,2} \leq \min(\pi_k^1, \pi_k^2)$ for all $k \in U$.

Frame

- transportation problem as method to solve PSC
- it enables us to compute the joint inclusion probability $p(s_1, s_2)$ of two samples drawn in two different occasions (s_1 and s_2) and the conditional probability $p(s_2|s_1)$. The latter enables us to choose the sample s_2 drawn in a second occasion given that the sample s_1 was drawn in the first occasion.

Transportation problem

The applications of the transportation problem in sample coordination are given in Raj (1968), Arthanari & Dodge (1981), Causey, Cox & Ernst (1985), Ernst & Ikeda (1992), Ernst (1996, 1998), Ernst & Paben (2002). We use the following form of the transportation problem presented in Causey, Cox & Ernst (1985):

$$\max \sum_{i=1}^m \sum_{j=1}^q c_{ij} p_{ij} \quad (1)$$

subject to the constraints

$$\left| \begin{array}{l} \sum_{j=1}^q p_{ij} = p_1(s_i^1), i = 1, \dots, m, \\ \sum_{i=1}^m p_{ij} = p_2(s_j^2), j = 1, \dots, q, \\ p_{ij} \geq 0, i = 1, \dots, m, j = 1, \dots, q, \end{array} \right.$$

with $c_{ij} = \#(s_i^1 \cap s_j^2)$, $p_1(s_i^1) = \text{prob}(s_i^1)$, $p_2(s_j^2) = \text{prob}(s_j^2)$, $p_{ij} = \text{prob}(s_i^1, s_j^2)$. $s_i^1 \in \mathcal{S}_1$ and $s_j^2 \in \mathcal{S}_2$ denote the possible samples in the first and second occasion, respectively, with $\#(\mathcal{S}_1) = m$ and $\#(\mathcal{S}_2) = q$. We suppose that $p_1(s_i^1) > 0$, $p_2(s_j^2) > 0$ in order to compute the conditional probabilities.

	s_1^2	s_2^2	s_3^2	...	s_q^2	Σ
s_1^1	$p(s_1^1, s_1^2)$	$p(s_1^1, s_2^2)$	$p_1(s_1^1)$
s_2^1	$p(s_2^1, s_1^2)$	$p(s_2^1, s_2^2)$	$p_1(s_2^1)$
s_3^1	$p_1(s_3^1)$
s_4^1	$p_1(s_4^1)$
s_5^1	$p_1(s_5^1)$
s_6^1	$p_1(s_6^1)$
s_7^1	$p_1(s_7^1)$
s_8^1	$p_1(s_8^1)$
s_9^1	$p_1(s_9^1)$
s_{10}^1	$p_1(s_{10}^1)$
...
s_m^1	$p_1(s_m^1)$
Σ	$p_2(s_1^2)$	$p_2(s_2^2)$	$p_2(s_3^2)$...	$p_2(s_q^2)$	1

Goals

- Give the conditions when the value of the objective function in the case of an optimal solution given by the transportation problem (the relative upper bound) is equal to the absolute upper bound = $\sum_{k \in U} \min(\pi_k^1, \pi_k^2)$ (generally, the relative upper bound is \leq the absolute upper bound).
- Develop a procedure to decide if the absolute upper bound can be reached or not. An algorithm based on the *Iterative Proportional Fitting* (IPF) procedure is used to give an optimal solution in the case of SC, without solving the linear program.

Maximal sample coordination

- it is the case where the absolute upper bound is equal to the relative upper bound

- A measure of positive coordination is the number of common sampled units in two occasions. Let n_{12} be this number. The goal is to maximize the expectation of n_{12} defined as

$$E(n_{12}) = \sum_{k \in U} \pi_k^{1,2} = \sum_{k \in U} \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} = \sum_{s_i^1 \in \mathcal{S}_1} \sum_{s_j^2 \in \mathcal{S}_2} \#(s_i^1 \cap s_j^2) p_{ij},$$

which is the objective function of problem (1).

- To maximize $E(n_{12})$ = to maximize the objective function of problem (1).
- In general, $\sum_{s_i^1 \in \mathcal{S}_1} \sum_{s_j^2 \in \mathcal{S}_2} \#(s_i^1 \cap s_j^2) p_{ij} \leq \sum_{k=1}^N \min(\pi_k^1, \pi_k^2)$.
- The absolute upper bound is reached when $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$, for all $k \in U$.

Proposition 1

We suppose that $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$, for all $k \in U$. The following conditions are fulfilled:

- a1. if $(\min(\pi_k^1, \pi_k^2) = \pi_k^1$ and $k \in s_i^1, k \notin s_j^2)$ then $p_{ij} = p(s_i^1, s_j^2) = 0$;
- b1. if $(\min(\pi_k^1, \pi_k^2) = \pi_k^2$ and $k \notin s_i^1, k \in s_j^2)$ then $p_{ij} = p(s_i^1, s_j^2) = 0$.

The reciprocal is also available.

Remark

The joint probabilities p_{ij} in problem (1) can be formulated as a matrix $\mathbf{P} = (p_{ij})_{m \times q}$. Proposition 1 enables us to set some p_{ij} to zero in order to find an optimal solution to problem (1).

Proposition 2

A feasible solution of problem (1) with the properties:

- a2. $p_{ij} = 0$ if there exists $k \in s_i^1, k \notin s_j^2$ and $\min(\pi_k^1, \pi_k^2) = \pi_k^1$;
- b2. $p_{ij} = 0$ if there exists $k \notin s_i^1, k \in s_j^2$ and $\min(\pi_k^1, \pi_k^2) = \pi_k^2$;

is an optimal solution.

Proposition 3

We note by $I = \{k \in U | \pi_k^1 \leq \pi_k^2\}$ the set of "increasing" units and by $D = \{k \in U | \pi_k^1 > \pi_k^2\}$ the set of "decreasing" units.

Suppose that all samples have the corresponding probabilities > 0 , and $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$, for all $k \in U$. Let $s_1 \in \mathcal{S}_1$. If at least one of the following relations is fulfilled for all $s_2 \in \mathcal{S}_2$:

- a3. $(s_1 \setminus s_2) \cap I \neq \emptyset$,
- b3. $(s_2 \setminus s_1) \cap D \neq \emptyset$,

then the two designs cannot be maximally coordinated. This proposition holds in the symmetric sense, too (if we change s_1 to s_2 , s_2 to s_1 , I to D and D to I).

Example 1

Let $U = \{1, 2, 3, 4\}$, $I = \{3, 4\}$ and $D = \{1, 2\}$. We draw in two distinct occasions samples of size 2 and 3, respectively. Below, we note the zero values given by using Proposition 1. By x we note the non-zero values. The sample $\{3, 4\}$ in the first occasion has on its row only the zero values. The maximal coordination is not possible, because $p_1(\{3, 4\}) \neq 0$. The same result is also available by using Proposition 3.

	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}
{1,2}	x	x	x	x
{1,3}	0	0	x	0
{1,4}	0	0	x	0
{2,3}	0	0	0	x
{2,4}	0	0	0	x
{3,4}	0	0	0	0

Impossible maximal coordination

The proposed algorithm

We propose the next algorithm based on the Propositions 1 and 2:

ep 1. Let $\mathbf{P} = (p_{ij})_{m \times q}$ be the matrix given by the independence between both designs:

$$p_{ij} = p_1(s_i^1)p_2(s_j^2), \text{ for all } i = 1, \dots, m, j = 1, \dots, q.$$

ep 2. Set the zeros to p_{ij} using Proposition 1.

ep 3. If the conditions of Proposition 3 are fulfilled, stop the algorithm and give the message "the absolute upper bound cannot be reached";
else apply the IPF procedure to restore the margins.

The correctness of the algorithm is assured by Proposition 2.

IPF procedure (Deming, Stephan, 1940)

Concerning the IPF procedure, in a first iteration indicated by the exponent (1) calculate for all rows $i = 1, \dots, m$

$$p_{ij}^{(1)} = p_{ij}^{(0)} \frac{p_1(s_i^1)}{p_1^{(0)}(s_i^1)}, \text{ for all } j = 1, \dots, q, \quad (2)$$

where $p_{ij}^{(0)} = p_1(s_i^1)p_2(s_j^2)$ and $p_1^{(0)}(s_i^1) = \sum_{j=1}^q p_{ij}^{(0)}$. Now the total rows $p_1(s_i^1)$ are satisfied. Calculate in a second iteration for all columns $j = 1, \dots, q$

$$p_{ij}^{(2)} = p_{ij}^{(1)} \frac{p_2(s_j^2)}{p_2^{(1)}(s_j^2)}, \text{ for all } i = 1, \dots, m, \quad (3)$$

where $p_2^{(1)}(s_j^2) = \sum_{i=1}^m p_{ij}^{(1)}$. Now the total columns $p_2(s_j^2)$ are satisfied. In a third iteration, the resulting $p_{ij}^{(2)}$ are used in recursion (2) for obtaining $p_{ij}^{(3)}$, and so on until convergence is attained.

Example 2

- Causey, Cox & Ernst (1985)
- The mathematical program gives the solution 0.88.
- Yet, $\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = 0.9$.
- Using Proposition 3 we observe that the samples $\{2, 5\}$ and $\{3, 5\}$ have in their rows only values equal to zero, and then the two designs cannot be maximally coordinated.

Matrix P in Example 2

	{1}	{2}	{3}	{4}	{5}		Σ
{1}	x	x	x	0	x		0.150
{2}	0	x	0	0	0		0.018
{3}	0	0	x	0	0		0.012
{4}	0	x	x	x	0		0.240
{5}	0	0	0	0	x		0.040
{1,4}	x	x	x	x	0		0.300
{1,5}	x	x	x	0	x		0.050
{2,4}	0	x	0	0	0		0.036
{2,5}	0	0	0	0	0		0.006
{3,4}	0	0	x	0	0		0.024
{3,5}	0	0	0	0	0		0.004
\emptyset	0	x	x	0	0		0.12
Σ	0.200	0.144	0.047	0.162	0.018		1

Example 2 - transformation

- We modify the example by letting $\pi_{\frac{1}{5}}^1 = 0.2$, instead of 0.1.
- Now, $I = \{2, 3\}$, $D = \{1, 4, 5\}$ and the samples in the first design have the corresponding probabilities:

0.1, 0.012, 0.008, 0.24, 0.08, 0.3, 0.1, 0.036, 0.012, 0.024, 0.008, 0.08.

- We apply the proposed algorithm on matrix \mathbf{P} . **The absolute upper bound is now reached.**

Matrix P after the application of the Steps 1 and 2

	{1}	{2}	{3}	{4}	{5}	Σ
{1}	0.04	0.015	0.005	0	0	0.0600
{2}	0	0.0018	0	0	0	0.0018
{3}	0	0	0.0004	0	0	0.0004
{4}	0	0.036	0.012	0.072	0	0.1200
{5}	0	0.012	0.004	0	0.008	0.0240
{1,4}	0.12	0.045	0.015	0.09	0	0.2700
{1,5}	0.04	0.015	0.005	0	0.01	0.0700
{2,4}	0	0.0054	0	0	0	0.0054
{2,5}	0	0.0018	0	0	0	0.0018
{3,4}	0	0	0.0012	0	0	0.0012
{3,5}	0	0	0.0004	0	0	0.0004
∅	0	0.012	0.004	0	0	0.0160
Σ	0.200	0.144	0.047	0.162	0.0180	1

The margins are not respected, because they are different from the true marginal probabilities.

Matrix P after the application of the Step 3

	{1}	{2}	{3}	{4}	{5}		Σ
{1}	0.098570	0.001287	0.000143	0	0		0.100
{2}	0	0.012	0	0	0		0.012
{3}	0	0	0.008	0	0		0.008
{4}	0	0.009583	0.001065	0.229352	0		0.240
{5}	0	0.003194	0.000355	0	0.076451		0.080
{1,4}	0.226073	0.002952	0.000328	0.070648	0		0.300
{1,5}	0.075358	0.000984	0.000109	0	0.023549		0.100
{2,4}	0	0.036	0	0	0		0.036
{2,5}	0	0.012	0	0	0		0.012
{3,4}	0	0	0.024	0	0		0.024
{3,5}	0	0	0.008	0	0		0.008
\emptyset	0	0.072	0.008	0	0		0.080
Σ	0.400	0.150	0.050	0.300	0.100		1

We get the good marginal probabilities.

Values of $c_{ij} = \#(s_i^1, s_j^2)$

	{1}	{2}	{3}	{4}	{5}
{1}	1	0	0	0	0
{2}	0	1	0	0	0
{3}	0	0	1	0	0
{4}	0	0	0	1	0
{5}	0	0	0	0	1
{1,4}	1	0	0	1	0
{1,5}	1	0	0	0	1
{2,4}	0	1	0	1	0
{2,5}	0	1	0	0	1
{3,4}	0	0	1	1	0
{3,5}	0	0	1	0	1
\emptyset	0	0	0	0	0

Minimal sample coordination

A similar algorithm can be constructed in the case of negative co-ordination, when the overlap is minimized. It is the case where

$$\max(0, \pi_k^1 + \pi_k^2 - 1) \leq \pi_k^{1,2}, \text{ for all } k \in U.$$

In an analogous way, the quantity $\sum_{k \in U} \max(0, \pi_k^1 + \pi_k^2 - 1)$ is called the **absolute lower bound**. Retaining the same constraints, we now seek to minimize the objective function of the problem (1). In general,

$$\sum_{i=1}^m \sum_{j=1}^q c_{ij} p_{ij} \geq \sum_{k \in U} \max(0, \pi_k^1 + \pi_k^2 - 1).$$

By setting $\max(0, \pi_k^1 + \pi_k^2 - 1) = \pi_k^{1,2}$, for all $k \in U$ a proposition similarly to Proposition 1 is given below.

Proposition 4

The absolute lower bound is reached iff the following conditions are fulfilled:

- a. if $(k \in s_i^1 \cap s_j^2 \text{ and } \pi_k^{1,2} = 0)$, then $p_{ij} = 0$,
- b. if $(k \notin s_i^1 \cup s_j^2 \text{ and } \pi_k^{1,2} = \pi_k^1 + \pi_k^2 - 1)$, then $p_{ij} = 0$,

for all $k \in U$.

Remark

The algorithm above can be applied in the case of minimal sample coordination by using Proposition 4 instead of Proposition 1, and the absolute lower bound instead of the absolute upper bound.

Conclusions

The drawback of using linear programming in sample coordination is its huge computational aspect. However, it is possible to construct an algorithm to compute the joint probability of two samples drawn on two different occasion, without solving a linear programm. The proposed algorithm is based on the Propositions 1 or 4 respectively, which identifies the conditions when the absolute upper bound or absolute lower bound respectively are reached and gives a modality to determine the joint sample probabilities equal to zero. The algorithm uses the IPF procedure, which assures a fast convergence. The algorithm has the complexity $O(m \times q \times \text{number of iterations in IPF procedure})$, which is low compared to the linear programm, and it is very easy to implement.