NEYMAN TYPE ALLOCATION THROUGH BASICS OF LINEAR ALGEBRA

to Neyman or not to Neyman...?

Jacek Wesołowski

Statistics Poland, Warsaw

2nd Congress of Polish Statistics
Warsaw, July 10-12, 2018

Plan

Stratified SRSWOR

A population U of $N = \# U$ units is stratified,

$$U = \bigcup_{h=1}^{H} U_h$$

with $N_h = \# U_h$, $h = 1, \ldots, H$.

For $h = 1, \ldots, H$, samples $\mathcal{S}_h$ of size n_h are drawn from U_h independently according to the SRSWOR.

The final sample is

$$\mathcal{S} = \bigcup_{h=1}^{H} \mathcal{S}_h.$$

π-estimator and its variance

Then

$$\hat{t}_{st} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in \mathcal{S}_h} y_k$$

is the π-estimator of the total $t = \sum_{k \in U} y_k$.

Its variance is

$$D^2_{st} = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} N_h S_h^2.$$

Neyman approach

Minimize D_{st}^2 as a function of $\underline{n} = (n_1, \ldots, n_H)$ under the constraint

$$n_1 + \ldots + n_H = n. \tag{1}$$

The Lagrange function is

$$F(n_1, \ldots, n_H; \lambda) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} + \lambda \sum_{h=1}^{H} n_h.$$

Then

$$\frac{\partial F}{\partial n_h} = -\frac{N_h^2 S_h^2}{n_h^2} + \lambda = 0.$$

Consequently,

$$n_h = \frac{N_h S_h}{\sqrt{\lambda}}.$$

By (1) we have $\sqrt{\lambda} = \frac{1}{n} \sum_{g=1}^{H} N_g S_g$.

Optimal allocation

Consequently,

$$n_h = n \frac{N_h S_h}{\sum_{g=1}^{H} N_g S_g}, \quad h = 1, \ldots, H. \tag{2}$$

Moreover,

$$D^2_{opt} = \frac{1}{n} \left(\sum_{h=1}^{H} N_h S_h \right)^2 - \sum_{h=1}^{H} N_h S_h^2.$$

The solution is positive only when

$$n < \frac{\left(\sum_{h=1}^{H} N_h S_h \right)^2}{\sum_{h=1}^{H} N_h S_h^2}.$$

Problems: we need $n_h \leq N_h$!

Let us write $n_h = N_h e^{-x_h^2} \leq N_h$. Then (1) assumes the form

$$\sum_{h=1}^{H} N_h e^{-x_h^2} = n \tag{3}$$

and the Lagrange function is

$$F(x_1, \ldots, x_h; \lambda) = \sum_{h=1}^{H} N_h S_h^2 e^{x_h^2} + \lambda \sum_{h=1}^{H} N_h e^{-x_h^2}.$$

Differentiation (after simplifications) gives

$$x_h(S_h^2 - \lambda\, e^{-2x_h^2}) = 0.$$

That is either $\mathbf{x_h = 0}$ or $\mathbf{e^{-x_h^2} = \frac{S_h}{\sqrt{\lambda}}}$.

A la Stenger & Gabler (2005), Metrika

There exists a maximal set of strata $\emptyset \subset \mathcal{H} \subset \{1, \ldots, H\}$ such that $n_h^* = N_h$, $h \in \mathcal{H}$.

Construction of $\mathcal{H}$:

1. Order strata according to: $S_1 \geq S_2 \geq \ldots \geq S_H$.
2. Let $\mathcal{H}_k = \{1, 2, \ldots, k-1\}$, $\mathcal{H}_0 = \emptyset$, $k = 1, \ldots, H$.

   Let $N_A = \sum_{h \in A} N_h$ for $A \subset \{1, \ldots, H\}$.
3. $\mathcal{H} = \mathcal{H}_{k^*}$, where

$$k^* = \min \left\{ k : \; (n - N_{\mathcal{H}_k}) \, \frac{S_k}{\sum_{h \notin \mathcal{H}_k} N_h S_h} < 1 \right\}.$$

Then

$$n_h^* = (n - N_{\mathcal{H}}) \frac{N_h S_h}{\sum_{g \notin \mathcal{H}} N_g S_g}, \quad h \notin \mathcal{H}.$$

Comparisons for $H = 66$, $n = 200000$, $N = 500000$

Algorithm *noptcond* by Gabler, Ganninger and Münnich (2012) and the new algorithm *SteGab*. Both implementations in R.

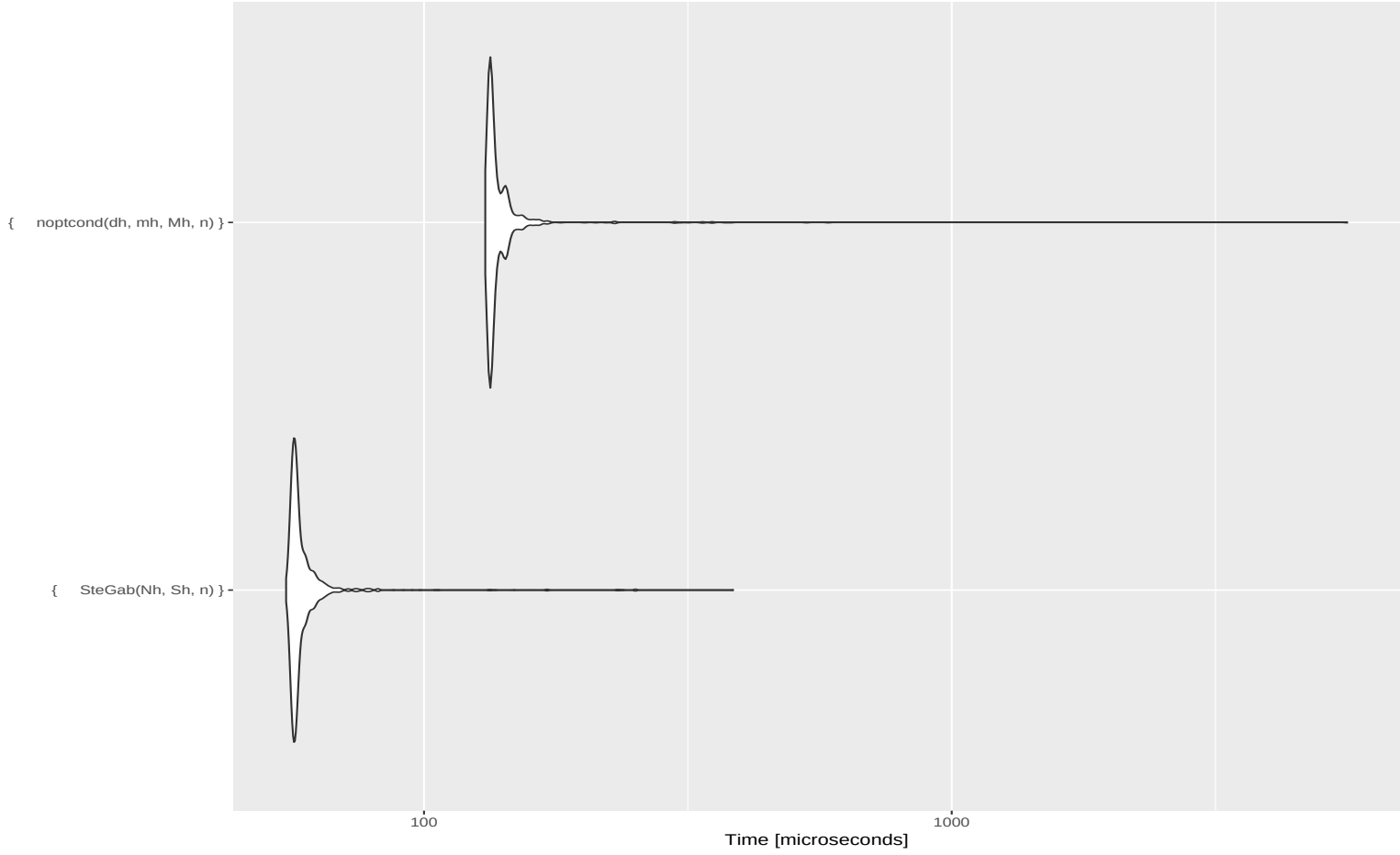


Figure: Time performace of *noptcond* and *SteGab* for 1000 runs.

Problems: we need n_h's to be positive integers!

A la Wright (2017): Since

$$\frac{1}{J} = 1 - \sum_{j=1}^{J-1} \frac{1}{k(k+1)}, \qquad J = 2, 3, \ldots,$$

$$D_{st}^2 = \sum_{h=1}^{H} N_h(N_h - 1)S_h^2 - \sum_{h=1}^{H} \sum_{k=1}^{n_h-1} \frac{N_h^2 S_h^2}{k(k+1)}.$$

Order strata according to

$$N_1 S_1 \geq N_2 S_2 \geq \ldots \geq N_H S_H.$$

Consider a rank-one matrix $A = [a_{h,k}]$ with $a_{h,k} = \frac{N_h^2 S_h^2}{k(k+1)}$.

Find n entries in the "upper left corner" of A, i.e. find n_h^* such that $N_h \geq n_h^* \geq n_{h+1}^*$ and such that

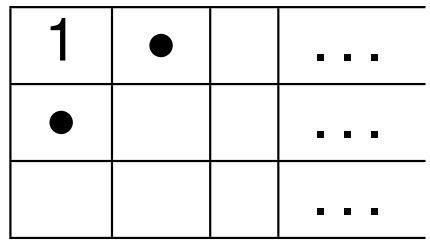
$$\sum_{h=1}^{H} \sum_{k=1}^{n_h^*-1} a_{h,k}$$

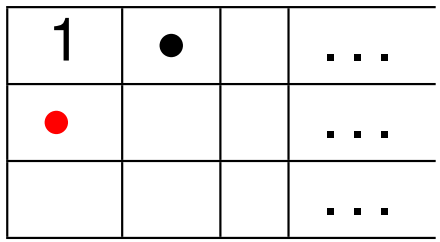
is maximal.

Integer partitions - a recursive algorithm

Any $n_1 \geq n_2 \geq \ldots \geq n_k \geq 1$ such that $n = n_1 + \ldots + n_K$ is called a size k integer partition of n.

A recursive algorithm starts with:

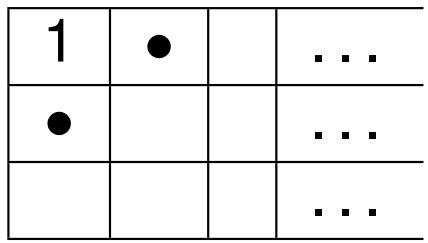


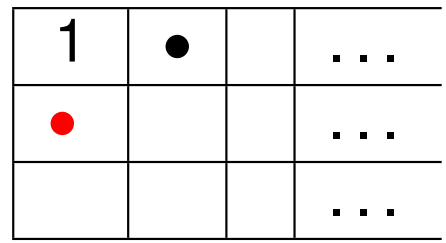


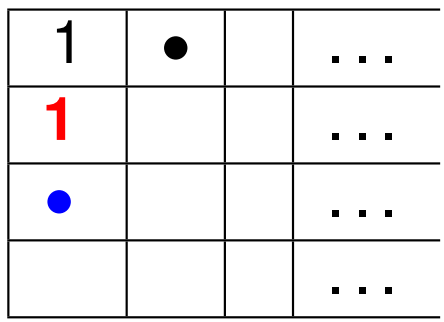
Integer partitions - a recursive algorithm

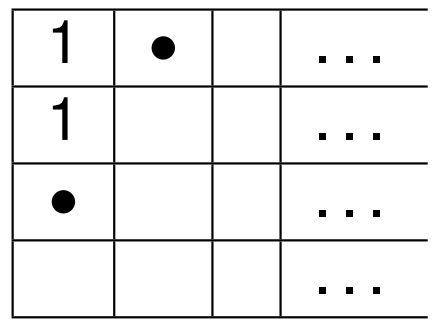
Any $n_1 \geq n_2 \geq \ldots \geq n_k \geq 1$ such that $n = n_1 + \ldots + n_K$ is called a size k integer partition of n.

A recursive algorithm starts with:







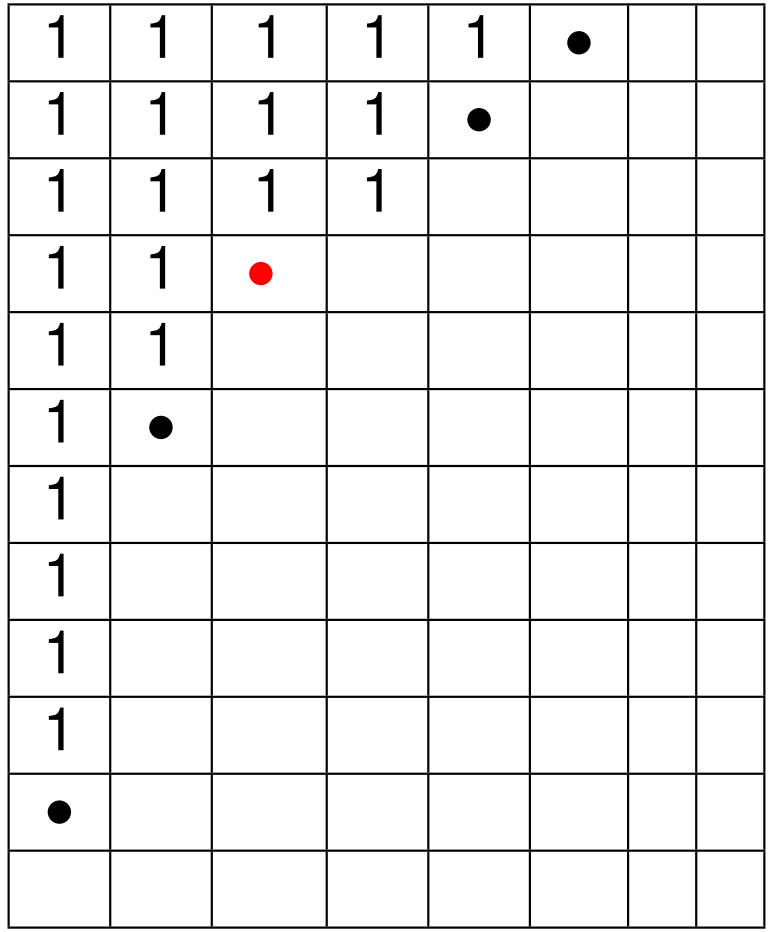


Recursive algorithm, cont.

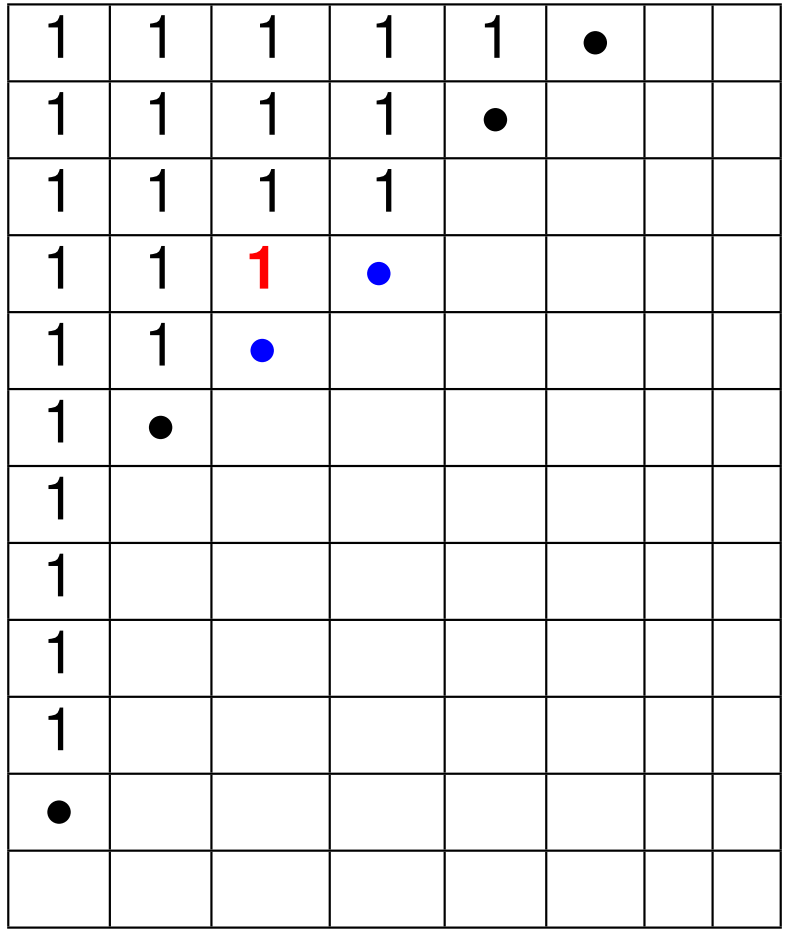
After 22 steps:

1	1	1	1	1	•		
1	1	1	1	•			
1	1	1	1				
1	1	•					
1	1						
1	•						
1							
1							
1							
1							
•							

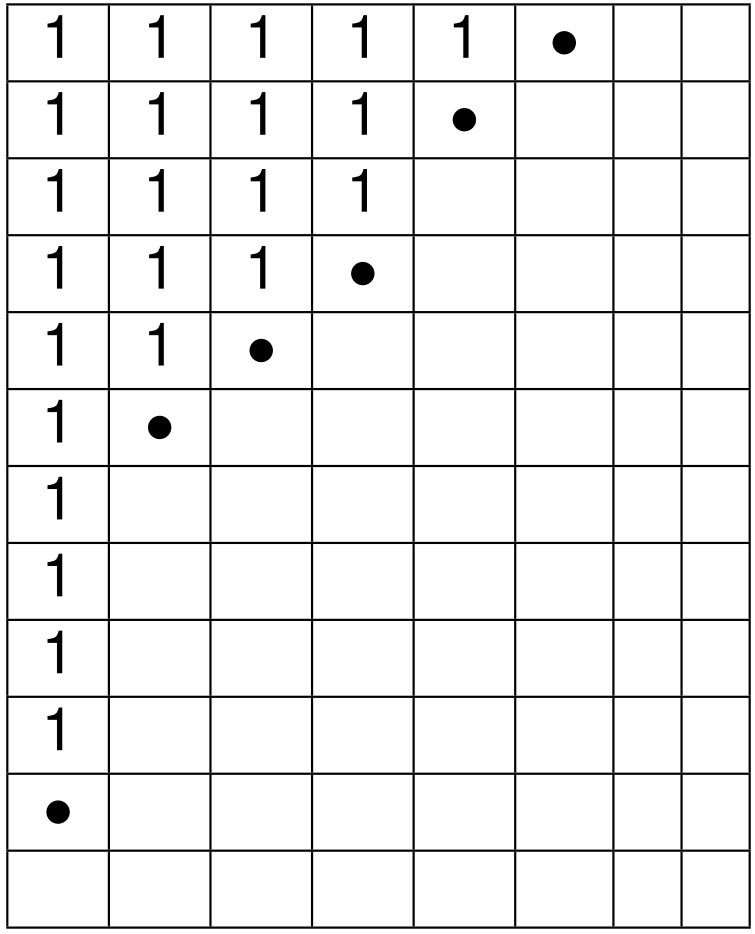
Recursive algorithm, cont.



Recursive algorithm, cont.



Recursive algorithm, cont.



Including box constraints

$$m_h \leq n_h \leq N_h, \quad h = 1, \ldots, H.$$

m_h	1	2	3	4	5	6	7	8	N_h
$m_1 = 3$	■	■	■	1	•		■	■	$N_1 = 6$
$m_2 = 1$	■	1	1	1			■	■	$N_2 = 6$
$m_3 = 2$	■	■	1	1		■	■	■	$N_3 = 5$
$m_4 = 4$	■	■	■	■				■	$N_4 = 7$
$m_5 = 1$	■	1	1	■	■	■	■	■	$N_5 = 3$
$m_6 = 1$	■	1	1	•				■	$N_6 = 7$
$m_7 = 1$	■	•						■	$N_7 = 7$
$m_8 = 1$	■			■	■	■	■	■	$N_8 = 3$
$m_9 = 2$	■	■						■	$N_9 = 7$
$m_{10} = 1$	■						■	■	$N_{10} = 6$

Choudhry, Rao, Hidiroglou (2012) - CRH

Let $\{U_i,\ i = 1, \ldots, I\}$, be a partition of U into (disjoint) domains in U. In each U_i a sample of n_i units is drawn by SRSWOR.

Minimize $n = n_1 + \ldots + n_I$ under the constraints:

$$\left(\frac{1}{n_i} - \frac{1}{N_i}\right) N_i^2 S_i^2 \leq RV_i, \quad i = 1, \ldots, I,$$

and

$$\sum_{i=1}^{I} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) N_i^2 S_i^2 \leq RV_0.$$

Solved in CRH by NLP.

See the monograph Valliant, Dever and Frauke (2013) for practical use of NLP to abundance of allocation issues.

CRH - an approach through integer partitions

Note that the domains variances constraints give

$$m_i := \lceil \frac{N_i^2 S_i^2}{RV_i + N_i S_i^2} \rceil \leq n_i \leq N_i, \quad i = 1, \ldots, I.$$

The allocation follows by choosing the smallest possible number of entries

$$n = n_1 + \ldots + n_I$$

in the matrix A except already chosen $\sum_{i=1}^{I} m_i$ entries such that

$$\sum_{i=1}^{I} \frac{N_i^2 S_i^2}{n_i} \leq \sum_{i=1}^{I} N_i S_i^2 + RV_0.$$

Stratification in the multi-domain setting

Every domain U_i is stratifed:

$$U_i = \bigcup_{h=1}^{H_i} U_{i,h}, \quad i = 1, \ldots, I.$$

In each domain stratified SRSWOR is used. The standard estimator of the total in U_i has variance

$$T_i = \sum_{h=1}^{H_i} \left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}} \right) N_{i,h}^2 S_{i,h}^2, \quad i = 1, \ldots, I.$$

The overall variance is $\sum_{i=1}^{I} T_i$.

Simulatneous minimization of domains variances and the overall variance

We introduce predesigned priority weights $\kappa_i > 0$, such that $\sum_{i=1}^{I} \kappa_i = 1$. and we write

$$T_i = \kappa_i V$$

(then V denotes *unknown* overall variance). Thus it suffices to minimize S under the constraints

$$\sum_{i=1}^{I} \sum_{h=1}^{H_i} n_{i,h} = n \tag{4}$$

and

$$\sum_{h=1}^{H_i} \left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}} \right) N_{i,h}^2 S_{i,h}^2 = \kappa_i V, \quad i = 1, \ldots, I. \tag{5}$$

Population matrix **D**

Let

$$\mathbf{D} = \frac{1}{n}\,\underline{a}\,\underline{a}^T - \mathrm{diag}(\underline{c})$$

be an $I \times I$ matrix with

$$\underline{a} = (a_1, \ldots, a_I)^T = \left(\frac{1}{\sqrt{\kappa_i}_i} \sum_{h=1}^{H_i} N_{i,h} S_{i,h},\ i = 1, \ldots, I \right)^T,$$

$$\underline{c} = (c_1, \ldots, c_I)^T = \left(\frac{1}{\kappa_i} \sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2,\ i = 1, \ldots, I \right)^T$$

and $\mathrm{diag}(\underline{c})$ is a diagonal matrix with $\underline{c}$ being its diagonal.

Note that **D** is a rank one perturbation of $\mathrm{diag}(\underline{c})$.

Neyman type allocation for domains

Theorem

Consider stratified SRSWOR in all domains (as described above) with the total sample size n. Assume that

$$n < \sum_{i=1}^{I} \frac{\left(\sum_{h=1}^{H_i} N_{i,h} S_{i,h}\right)^2}{\sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2}. \tag{6}$$

Then there exists the unique, simple and positive eigenvalue λ^* *of matrix* **D** *and* $\underline{v}^* = (v_1^*, \ldots, v_I^*)$ *the respective unit eigenvector with all entries positive.*

Neyman type allocation for domains

Theorem (cont)

The multi-domain optimal allocation (with priority weights κ_i*,* $i = 1, \ldots, I$*), i.e. with the minimal* V *under the sample size constraints* (4) *and* (5) *(given above) has the form*

$$n_{i,h} = n \frac{v_i^* N_{i,h} S_{i,h} / \sqrt{\kappa_i}}{\sum_{r=1}^{I} v_r^* \sum_{g=1}^{H_r} N_{r,g} S_{r,g} / \sqrt{\kappa_r}}. \tag{7}$$

Moreover,

$$V_{opt} = \frac{\left(\sum_{i=1}^{I} \frac{\sqrt{\kappa_i}}{v_i^*} \sum_{h=1}^{H_i} N_{i,h} S_{i,h} \right) \left(\sum_{i=1}^{I} \frac{v_i^*}{\sqrt{\kappa_i}} \sum_{h=1}^{H_i} N_{i,h} S_{i,h} \right)}{n} - \sum_{i=1}^{I} \sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2.$$

Special cases: no domains or no strata

If $I = 1$, i.e. when there are **no domains**, the above formulas are reduced to the classical Neyman ones.

If $H_i = 1$, $i = 1, \ldots, I$, i.e. when **domains are not stratified**, an alternative solution to that given by CRH is obtained:

Corollary

The optimal domain-wise efficient allocation has the form

$$n_i = n \frac{v_i^* N_i S_i / \sqrt{\kappa_i}}{\sum_{j=1}^{I} v_j^* N_j S_j / \sqrt{\kappa_j}}, \quad i = 1, \ldots, I.$$

Moreover,

$$V_{opt} = \left[\frac{1}{n} \left(\sum_{i=1}^{I} \frac{\sqrt{\kappa_i}}{v_i^*} N_i S_i \right) \left(\sum_{i=1}^{I} \frac{v_i^*}{\sqrt{\kappa_i}} N_i S_i \right) - \sum_{i=1}^{I} N_i S_i^2 \right].$$

Stratification on the first and the second stage

A collection $\mathcal{V}_i$ of PSUs from ith domain of U is stratified: $\mathcal{V}_i = \bigcup_{h=1}^{H_i} \mathcal{V}_{i,h}$ and $M_{i,h} = \#\,\mathcal{V}_{i,h}$.

A collection of SSUs in jth PSU from $\mathcal{V}_{i,h}$ is stratified into $\bigcup_{g=1}^{G_{j,h,i}} \mathcal{W}_{i,h,j,g}$ and $N_{i,h,j,g} = \#\,\mathcal{W}_{i,h,j,g}$.

Let

$$D^2_{i,h} = \frac{1}{M_{i,h}-1} \sum_{j \in \mathcal{V}_{i,h}} \left(t_j - \bar{t}_{i,h} \right)^2 ,$$

$$\gamma_{i,h} = M_{i,h} \left(M_{i,h} D^2_{i,h} - \sum_{j \in \mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} N_{i,h,j,g}\, S^2_{i,h,j,g} \right) ,$$

$$\beta_{i,h,j,g} = N_{i,h,j,g}\, S_{i,h,j,g}.$$

Single constraint on the overall cost

Theorem

Consider the constraint for the expected variable cost

$$\sum_{i=1}^{I}\sum_{h=1}^{H_i} c^2_{I,i,h} m_{i,h} + \sum_{i=1}^{I}\sum_{h=1}^{H_i} \frac{m_{i,h}}{M_{i,h}} \sum_{j\in\mathcal{V}_{i,h}} c^2_{II,i,h,j} \sum_{g=1}^{G_{i,h,j}} n_{i,h,j,g} = C \quad (8)$$

and constraints given by relative priorities of domain variances

$$\sum_{h=1}^{H_i} \frac{\gamma_{i,h} + M_{i,h} \sum_{j\in\mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \frac{\beta^2_{i,h,j,g}}{n_{i,h,j,g}}}{m_{i,h}} - \sum_{h=1}^{H_i} M_{i,h} D^2_{i,h} = \kappa_i V. \quad (9)$$

Theorem (cont.)

Assume that $\gamma_{i,h} > 0$ *for all* $h = 1, \ldots, H_i$, $i = 1, \ldots, I$.

Let

$$\mathbf{D} = \frac{\underline{a}\,\underline{a}^T}{C} - \mathrm{diag}(\underline{c}),$$

where $\underline{a} = (a_i,\ i = 1, \ldots, I)^T$, $\underline{c} = (c_i,\ i = 1, \ldots, I)$,

$$a_i = \frac{\nu_i}{\sqrt{\kappa_i}}, \quad \textit{with} \quad \nu_i = \sum_{h=1}^{H_i} \left(c_{I,i,h} \sqrt{\gamma_{i,h}} + \sum_{j \in \mathcal{V}_{i,h}} c_{II,i,h,j} \sum_{g=1}^{G_{i,h,j}} \beta_{i,h,j,g} \right),$$

and

$$c_i = \frac{1}{\kappa_i} \sum_{h=1}^{H_i} M_{i,h}\, D_{i,h}^2, \quad i = 1, \ldots, I.$$

Theorem (cont.)

Assume that

$$\sum_{i=1}^{I} \frac{\nu_i}{\sum_{h=1}^{H_i} M_{i,h} D_{i,h}^2} > C.$$

Then **D** *has a unique, positive simple eigenvalue* λ^* *with a respective unit eigenvector* $\underline{v}^* = (v_1^*, \ldots, v_I^*)^T$ *having all entries positive.*

(the proof uses here the Perron-Frobenius theorem)

Theorem (cont.)

The multi-domain optimal allocation is given by

$$m_{i,h} = C \frac{v_i^* \sqrt{\gamma_{i,h}}}{\sqrt{\kappa_i} c_{I,i,h} \sum_{r=1}^{I} v_r^* \nu_r / \sqrt{\kappa_r}}$$

and

$$n_{i,h,j,g} = \frac{c_{I,i,h} M_{i,h} \beta_{i,h,j,g}}{c_{II,i,h,j} \sqrt{\gamma_{i,h}}}.$$

Moreover, the minimal variances in the domains are

$$T_{i,opt} = \kappa_i V_{opt}, \qquad i = 1, \ldots, I,$$

with

$$V_{opt} = \lambda^* = \frac{1}{C} \left(\sum_{i=1}^{I} \frac{\sqrt{\kappa_i}}{v_i^*} \nu_i \right) \left(\sum_{i=1}^{I} \frac{v_i^*}{\sqrt{\kappa_i}} \nu_i \right) - \sum_{i=1}^{I} \sum_{h=1}^{H_i} M_{i,h} D_{i,h}^2$$

being the optimal overall variance.

Separate constraints on expected sizes of PSU and SSU samples

PSUs sample size constraint:

$$\sum_{i=1}^{I} \sum_{h=1}^{H_i} m_{i,h} = m \tag{10}$$

SSUs expected sample size constraint:

$$\sum_{i=1}^{I} \sum_{h=1}^{H_i} \frac{m_{i,h}}{M_{i,h}} \sum_{j \in \mathcal{W}_{i,h}} \sum_{g=1}^{G_{i,h,j}} n_{i,h,j,g}. \tag{11}$$

D as rank-two perturbation of a diagonal matrix

Let

$$\mathbf{D} = \frac{\underline{a}\,\underline{a}^T}{m} + \frac{\underline{b}\,\underline{b}^T}{n} - \mathrm{diag}(\underline{c}),$$

where $\underline{a}$ and $\underline{c}$ are as above and $\underline{b} = (b_i,\ i = 1, \ldots, I)^T$ is defined by

$$b_i = \frac{1}{\sqrt{\kappa_i}} \sum_{h=1}^{H_i} \sum_{j \in \mathcal{W}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \beta_{i,h,j,g}, \quad i = 1, \ldots, I.$$

Lemma

If **D** *has a positive eigenvalue* λ^* *then it is simple and unique and respective unit eigenvector* $\underline{v}^* = (v_1^*, \ldots, v_I^*)$ *can be chosen in such a way that* $v_i^* > 0$, $i = 1, \ldots, I$.

Theorem

Under the sample sizes constraints (10) *and* (11) *and under*

$$\sum_{h=1}^{H_i} \frac{1}{m_{i,h}} \left(\gamma_{i,h} + M_{i,h} \sum_{j \in \mathcal{V}_{i,h}} \sum_{g=1}^{G_{i,h,j}} \frac{\beta_{i,h,j,g}^2}{n_{i,h,j,g}} \right) - \sum_{h=1}^{H_i} M_{i,h} D_{i,h}^2 = \kappa_i V,$$

the optimal allocation is

$$m_{i,h} = m \frac{v_i^* \sqrt{\gamma_{i,h}}}{\sum_{r=1}^{I} v_r^* \sum_{k=1}^{H_r} \sqrt{\gamma_{r,k}}}.$$

and

$$n_{i,h,j,g} = n \frac{v_i^* M_{i,h} \beta_{i,h,j,g}}{m_{i,h} \sum_{r=1}^{I} v_r^* \sum_{k=1}^{H_r} \sum_{s \in \mathcal{W}_{r,k,s}} \sum_{\ell=1}^{G_{r,k,s}} \beta_{r,k,s,\ell}}.$$

Summary of allocation methodology

- classical Neyman (1934); also Tchuprow (1923);
- NLP approach (*black box* style) e.g. Valliant, Dever & Frauke (2013)
- box constraints *a la* Stenger & Gabler (2005) and Gabler, Ganninger & Münich (2012);
- both, box constraints and integer solution *a la* Wright (2017);
- eigenproblem approach in multi-domain and multi-stage case: Niemiro & JW (2001), Kozak, Zieliński & Singh (2008), JW & Wieczorkowski (2017), Khan & JW (2018), JW (2018+).

Literature

- Choudhry, G.H., Rao, J.N.K., Hidiroglou, M.A., On sample allocation for efficient domain estimation. *Survey Meth.* **38(1)** (2012), 23-29.

- Khan, M.G.M., Wesołowski, J. , Neyman-type sample allocation for domains efficient estimation in multi-stage sampling. *Appl. Statist. Anal.* (2018) - under "minor revision"

- Niemiro, W., Wesołowski, J., Fixed precision allocation in two-stage sampling. *Appl. Math.* **28** (2001), 73-82.

- Stenger, H., Gabler, S., Combining random sampling and census strategies - Justification of inclusion probabilities equal to 1. *Metrika* **61** (2005), 137-156.

- Valliant, R., Dever, J.A., Frauke, K., *Practical Tools for Designing and Weighting Sample Surveys*, Springer, 2013.

- Wesołowski, J. , Multi-domain Neyman optimal allocation. *manuscript* (2018) - under preparation.

- Wesołowski, J., Wieczorkowski, R., An eigenproblem approach to optimal equal-precision sample allocation in subpopulations. *Comm. Statist. Theory Meth.* **46(5)** (2017), 2212-2231.