Surveying populations when individual sampling costs are not homogenous

Wojciech Gamrot

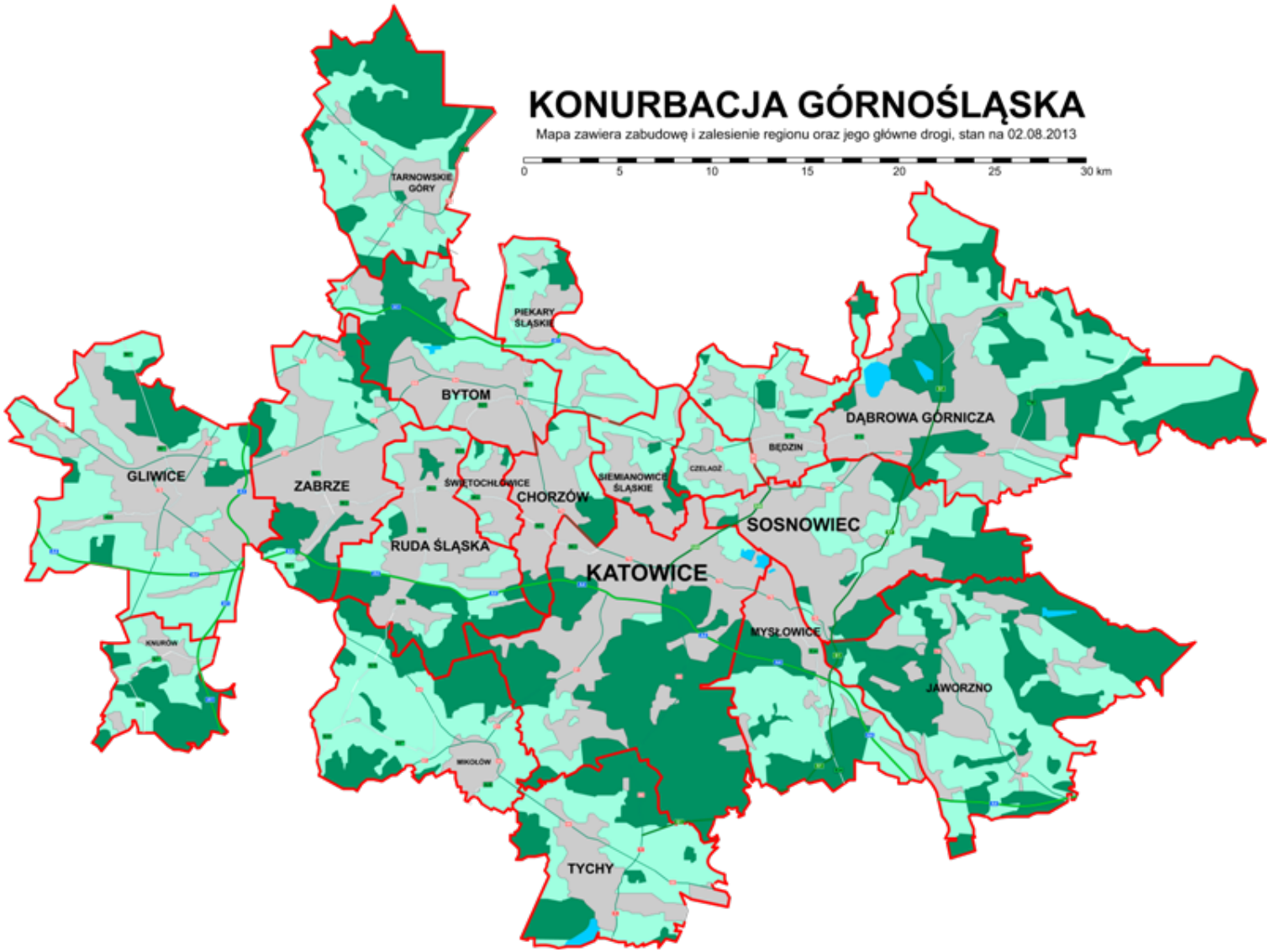
Department of Statistics Econometrics and Mathematics
University of Economics in Katowice

Outline

- Cost model
- Sample cost under SRSWOR
- Simple random sampling
- Pathak scheme
- Greedy scheme
- Simulation
- Variance estimation

A motivating example
Poland (by night - according to NASA)

A closer look at that big spot



The center of Katowice

A typical silesian bus

Need this to travel by mass transit

A ticket issued by Municipal Transport Union of Upper Silesia (KZK-GOP)

Examples of varying-cost situations

- Sampling communication lines
- Sampling in auditing
- Spatial dispersion (forest □'s, companies, households)
- Costs ~ unit size (buildings, cultivation, ecologic assets)
- Varying scope of required data (fiscal issues, medicine)
- Cluster sampling / Two-stage sampling
- Non-respondents to be followed up
- Recursive / adaptive sampling

Cost is a useful abstraction

It may include more than money

Other resources may be limited as well

- Time
- Equipment
- Administrative permissions

Here, we shall concentrate on a single resource

Notation

- Finite population $U = \{1, ..., N\}$
- Fixed characteristic $y_1, ..., y_N$
- Parameter under study $\overline{y} = N^{-1} \sum_{i \in U} y_i$
- Sample $s \subset U$
- Sampling design $P(s)$
- 1st-order inclusion probability: $\pi_i = Pr(i \in s)$
- 2nd-order inclusion probability: $\pi_{ij} = Pr(i, j \in s)$

Sample cost

A general definition

The cost C_s of the sample s is the amount of resources spent to observe desired characteristics of units in s

Assumed model (Skibicki & Wywial 2002,2003)

1. Individual costs $c_1, ..., c_N$ are associated with units in U
2. $C_s = \sum_{i \in s} c_i$
3. $c_1, ..., c_N$ are constant
4. $c_1, ..., c_N$ are known in advance

Sample cost

Specific model features

- Finite population assumed (vs renewal processes)
- Individual costs differ (vs most of literature)
- Individual costs non-random (vs Khan models)
- Individual costs known in advance (vs Pathak/Kremers)
- Sample cost C_s generally random
- Distribution of C_s depends on $P(s)$
- It may be assessed in advance

SRSWOR
a special case

- Fixed sample size n
- 1st order inclusion probability: $\pi_i = \frac{n}{N}$
- 2nd order inclusion probability: $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
- Sample space given n: $\Omega_n = \{s \subset U : \#s = n\}$
- Size of Ω_n: $\frac{N!}{n!(N-n)!}$
- Distribution of C_s: $Pr\{C_s = a\} = \sum_{s:C_s=a} P(s)$

A little more notation

- Assume $c_1 \leq ... \leq c_N$ (no loss of generality)
- Let

$$\bar{c} = \frac{1}{N} \sum_{i \in U} c_i \tag{1}$$

$$S_U^2(c) = \frac{1}{N-1} \sum_{i \in U} (c_i - \bar{c})^2 \tag{2}$$

C_s distribution characteristics:

- expectation $E(C_s) = n\overline{c}$
- variance $V(C_s) = n\left(1 - \frac{n}{N}\right) S_U^2(c)$
- minimum $c_{min}(n) = \sum_{i=1}^{n} c_i$
- maximum $c_{max}(n) = \sum_{i=1}^{n} c_{N-i+1}$

More notation

- Survey budget: L
- Budget-use coefficient:

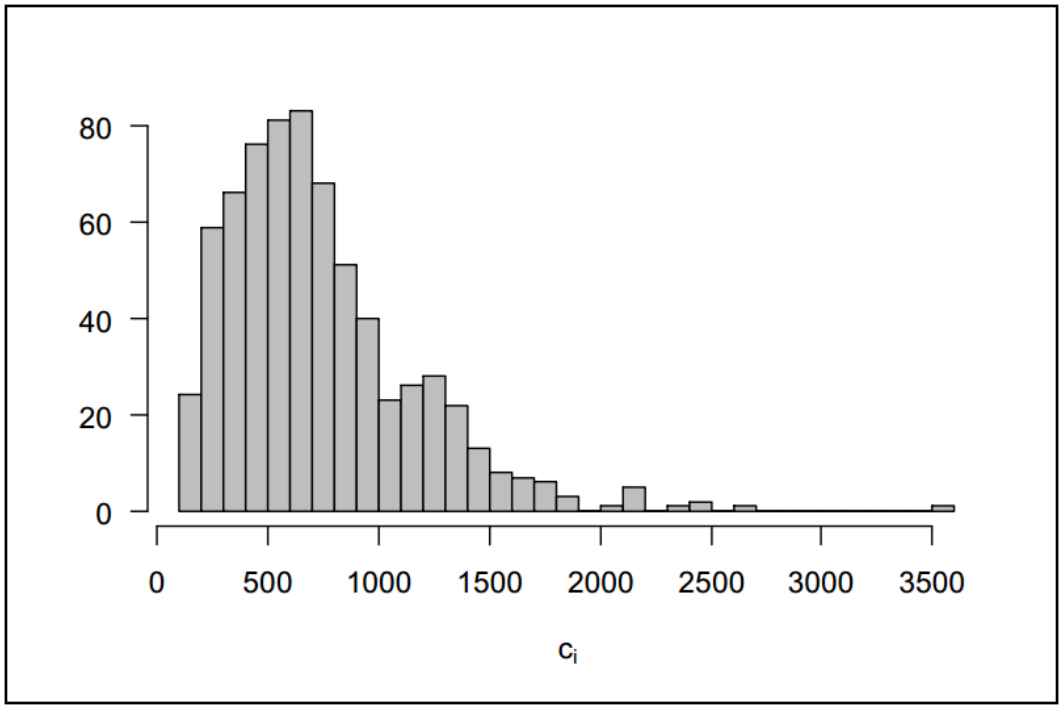
$$W_s = \frac{C_s}{L} \tag{3}$$

- If $c_1 = ... = c_N = \overline{c}$ and $L = z\overline{c}$ for $z \in \mathcal{Z}$ then $L = 1$
- In general:

$$\frac{c_{min}(n)}{L} = w_{min}(n) \leq W_s \leq w_{max}(n) = \frac{c_{max}(n)}{L} \tag{4}$$

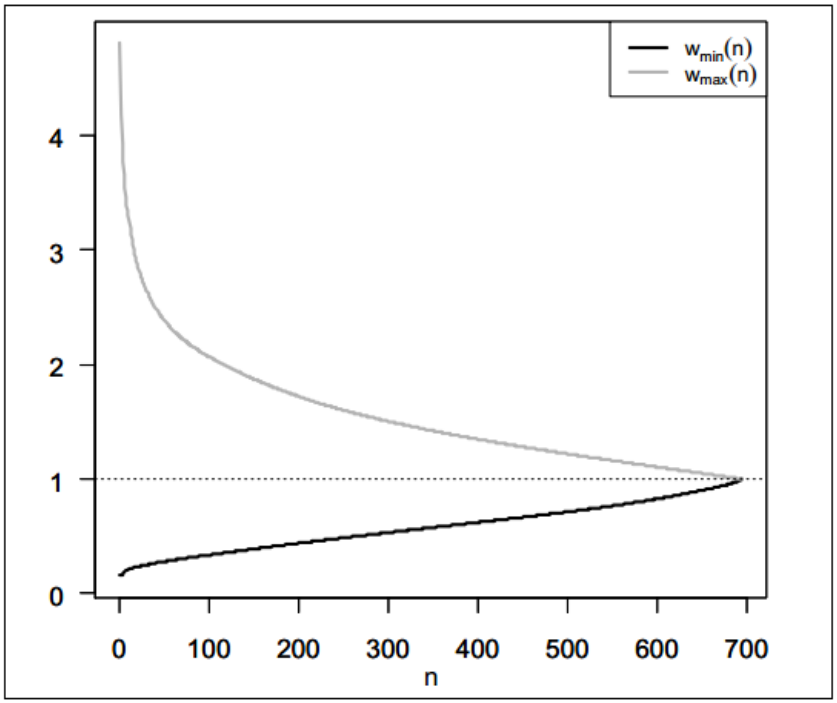
Example 1

- Population of 695 farms in Dabrowa Tarnowska district
- Sampling cost proportional to farm area:



Example 1

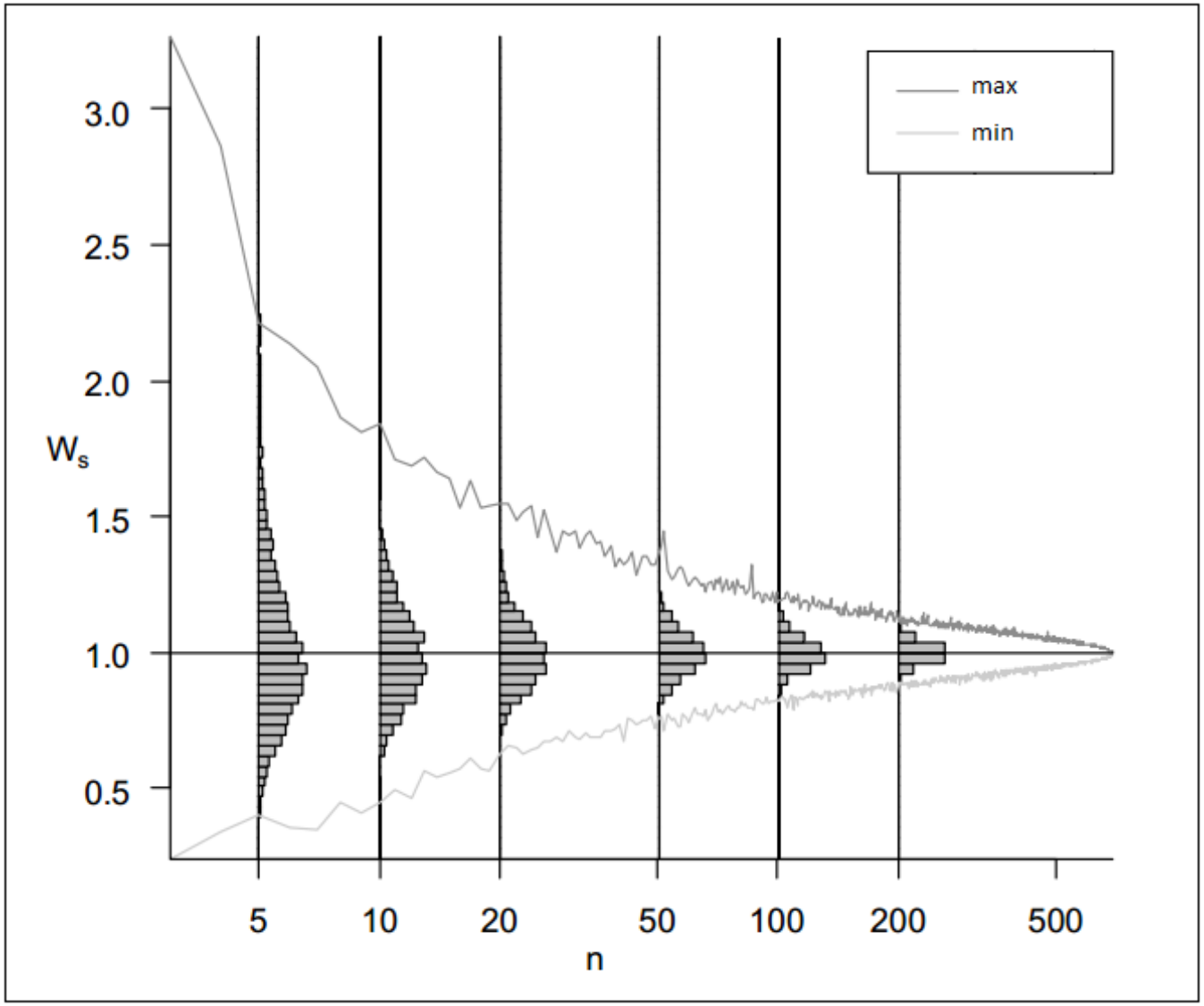
- False assumption: $c_1 = \ldots = c_N = \overline{c}$
- Sample size rule: $n = \lfloor \frac{L}{\overline{c}} \rfloor$
- Simplifying to $L = n\overline{c}$ we get:



Example 1
numbers

n	$w_{\min}(n)$	$w_{\max}(n)$
100	0.34	2.07
200	0.44	1.72
300	0.53	1.50
400	0.62	1.35
500	0.71	1.22
600	0.83	1.11

Example 2
A simulation study for farm area data



Example 2

Observed frequencies:

n	Deviations >10% in plus	Deviations >10% in minus	Total
50	0.1058	0.1025	0.2083
100	0.0350	0.0255	0.0605

For other populations it may be worse.

Interim conclusion

If nothing is done:

- deviations in plus:
  - generate financial losses
  - may render the survey unfeasible
  - may force reduction of the sample
  - lead to 'unofficial' rejective sampling (and biases)
- deviations in minus:
  - may trigger funding cuts for next edition of the survey
  - undermine credibility of survey results
  - suggest that unused funds may improve accuracy

Budget margin
An ad-hoc solution

- Budget never exceeded for

$$n_x = max\{n : c_{max}(n) \leq L\} \tag{5}$$

- Equivalent to a margin of funds:

$$\Delta = L - n_x \overline{c} \tag{6}$$

- The way to go ?

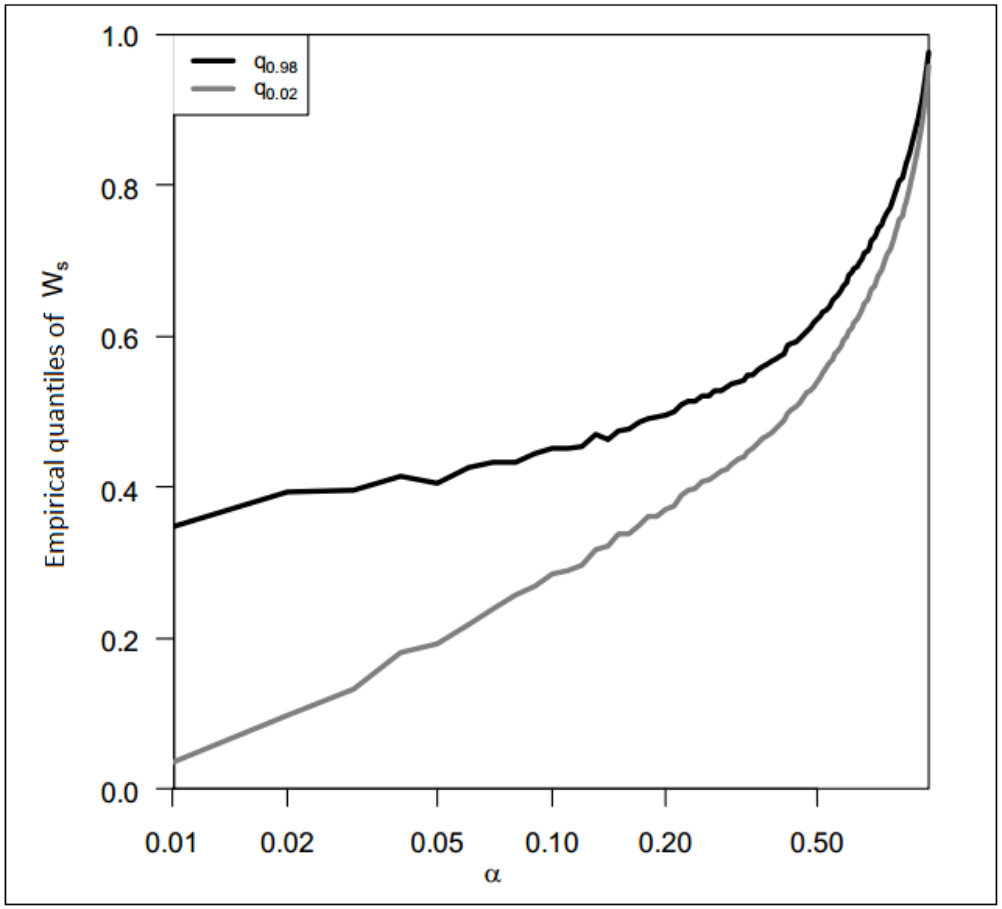
Example 3
Farm data again

Calculations for $L = \alpha(c_1 + ... + c_N)$:

α	n	n_x	α	n	n_x
0,01	6	1	0,15	104	42
0,02	13	3	0,20	139	60
0,03	20	5	0,30	208	101
0,04	27	8	0,40	278	147
0,05	34	10	0,50	347	202
0,06	41	13	0,60	417	265
0,07	48	16	0,70	486	337
0,08	55	19	0,80	556	421
0,09	62	22	0,90	625	525
0,10	69	25	1,00	695	695

Example 3
Empirical distribution of W_s

Quantiles of W_s:



For $\alpha \leq 0.2$ less than half budget spent in 98% of samples !

Pathak (1976) fixed-cost sampling

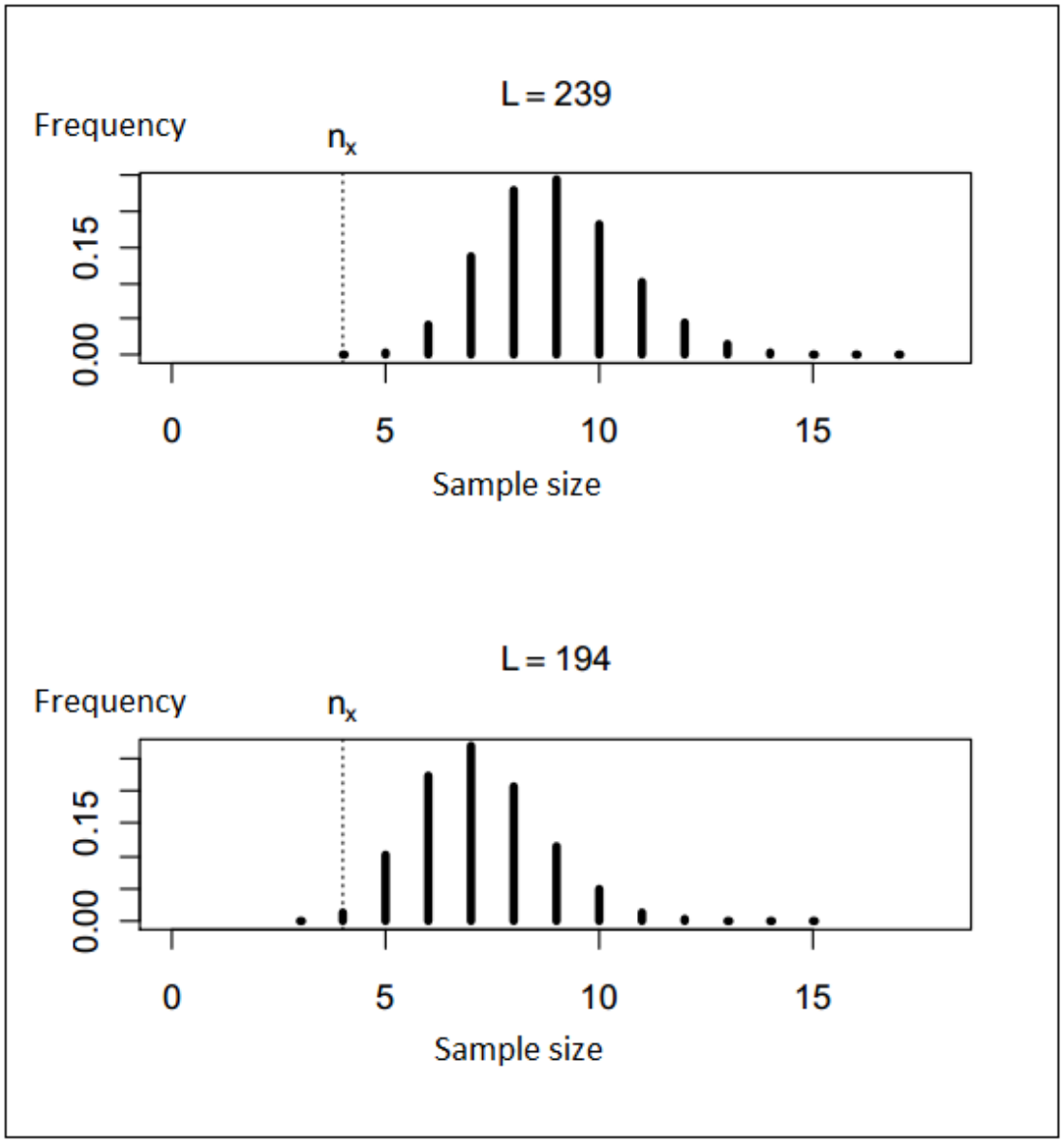
- Units are drawn
  - one by one
  - with equal probabilities
  - without replacement
  - while the cumulative cost is lower than L
  - the unit which breaks this is not included, (at no cost)
- Sample $(A_1, ..., A_M)$

Pathak fixed-cost sampling
features

- Budget never exceeded
- Random sample cost
- Random sample size M
- min(M) and max(M) easily calculated

Example 4

$N = 50$, $\mathbf{c} = [1, 2, ..., 50]'$, Simulation: 10^6 samples



Pathak fixed-cost sampling
inclusion probabilities

First-order inclusion probability

$$\pi_i = \frac{\Phi_i}{N!}$$

Second-order inclusion probability

$$\pi_{ij} = \frac{\Phi_{ij}}{N!}$$

where:

Φ_i - # permutations of U resulting in drawing the i-th unit

Φ_{ij} - # permutations of U resulting in drawing of (i, j)

Dependent on costs and usually hard to calculate.

Estimation

- Values $Y_1, ..., Y_M$ of study variable observed
- Inclusion probabilities not known
- Computation of Horvitz-Thompson estimator problematic
- Pathak's (1976) unbiased estimator:

$$\overline{Y}_M = \frac{1}{M} \sum_{i=1}^{M} Y_i \tag{7}$$

Estimation

- Its variance:

$$V(\overline{Y}_M) = \frac{1}{2N(N-1)} \sum_{i \neq j \in U} \left(\Lambda_{ij} - \frac{1}{N} \right) (y_i - y_j)^2 \qquad (8)$$

where

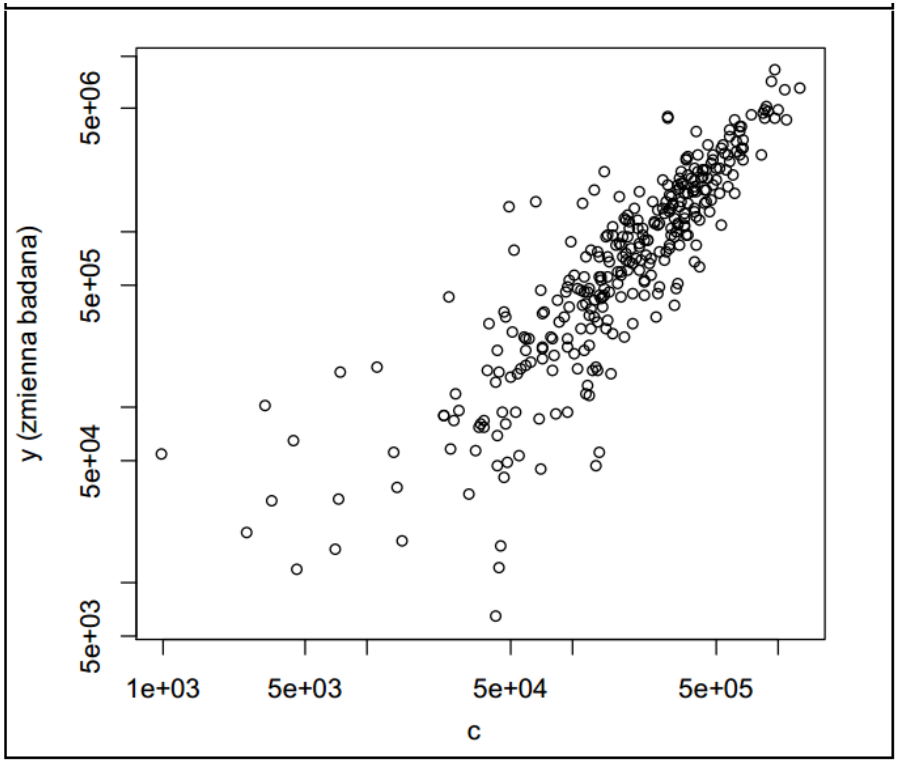
$$\Lambda_{ij} = E\left(\frac{1}{M} \middle| A_1 = i, A_2 = j \right) \qquad (9)$$

- Variance estimator:

$$\hat{V}(\overline{Y}_M) = \left(\frac{1}{M} - \frac{1}{N} \right) \frac{1}{2M(M-1)} \sum_{i \neq j=1}^{M} (Y_i - Y_j)^2 \qquad (10)$$

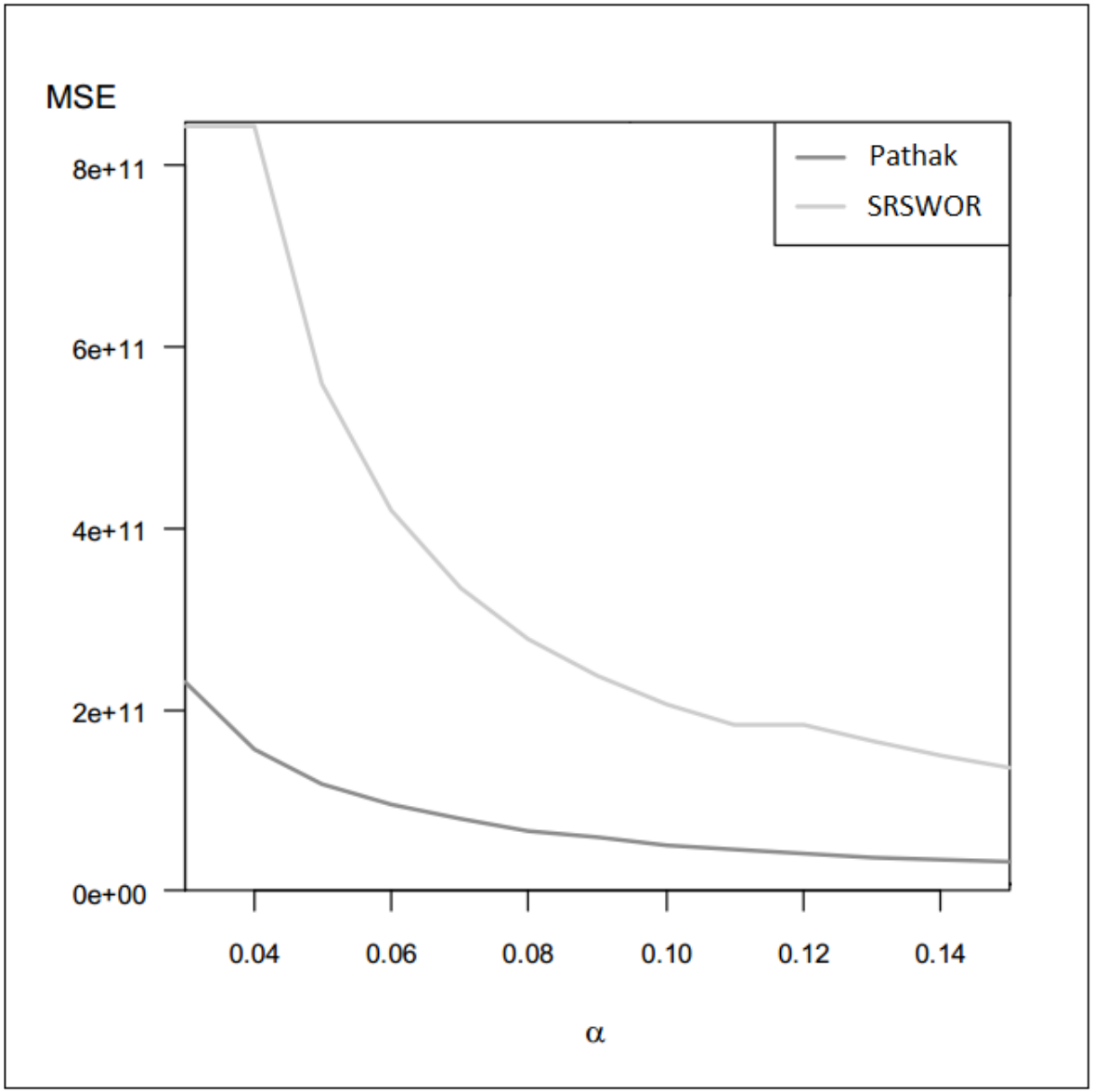
Example 5

- A population of $N = 319$ KZK-GOP communication lines
- Study variable: yearly number of passengers
- Sampling cost proportional to yearly vehicle miles driven
- Joint distribution of cost vs study variable



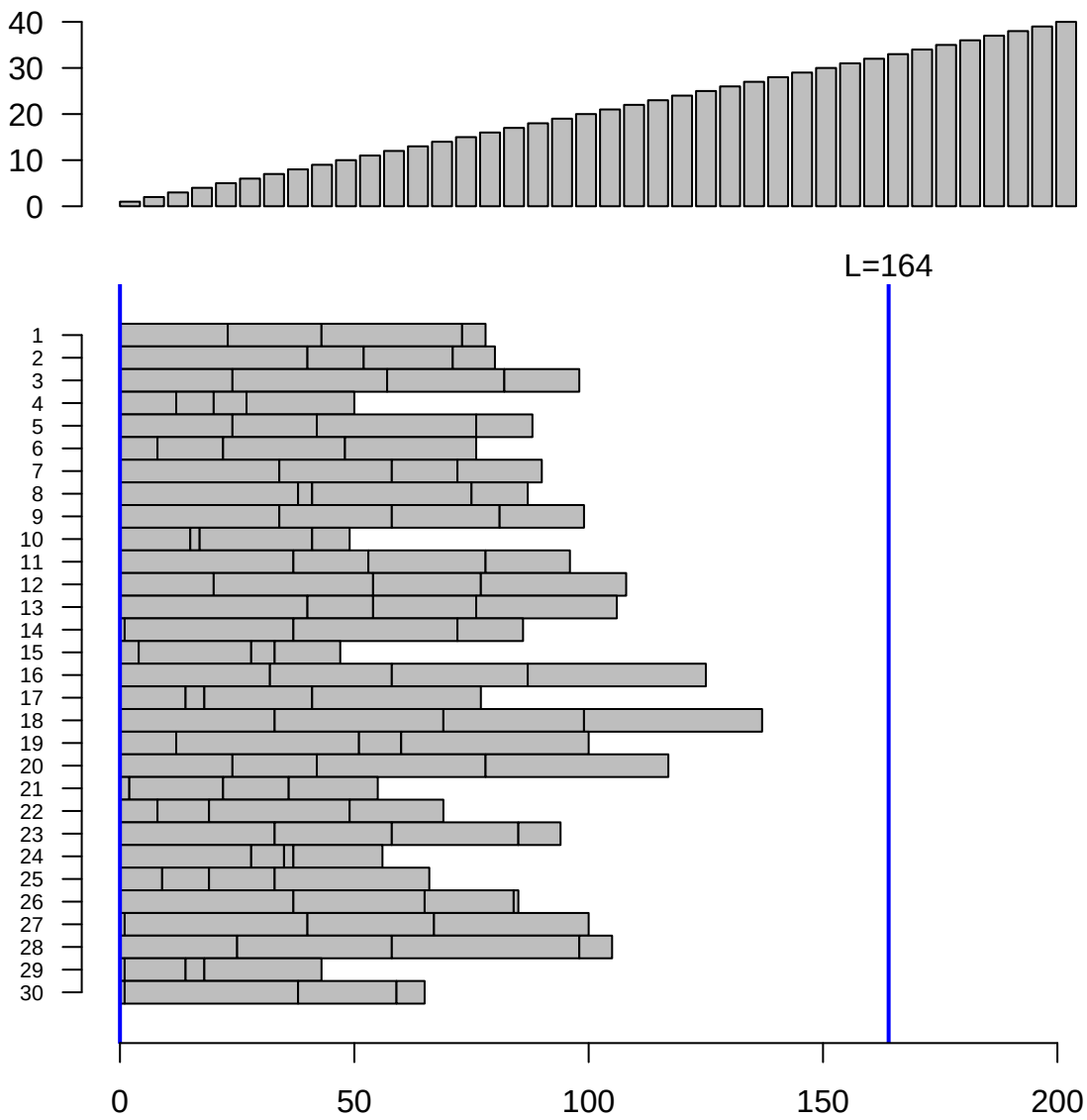
Example 5

Estimator accuracy for various budgets ($\alpha = L/(c_1 + ... + c_N)$ (Simulation)



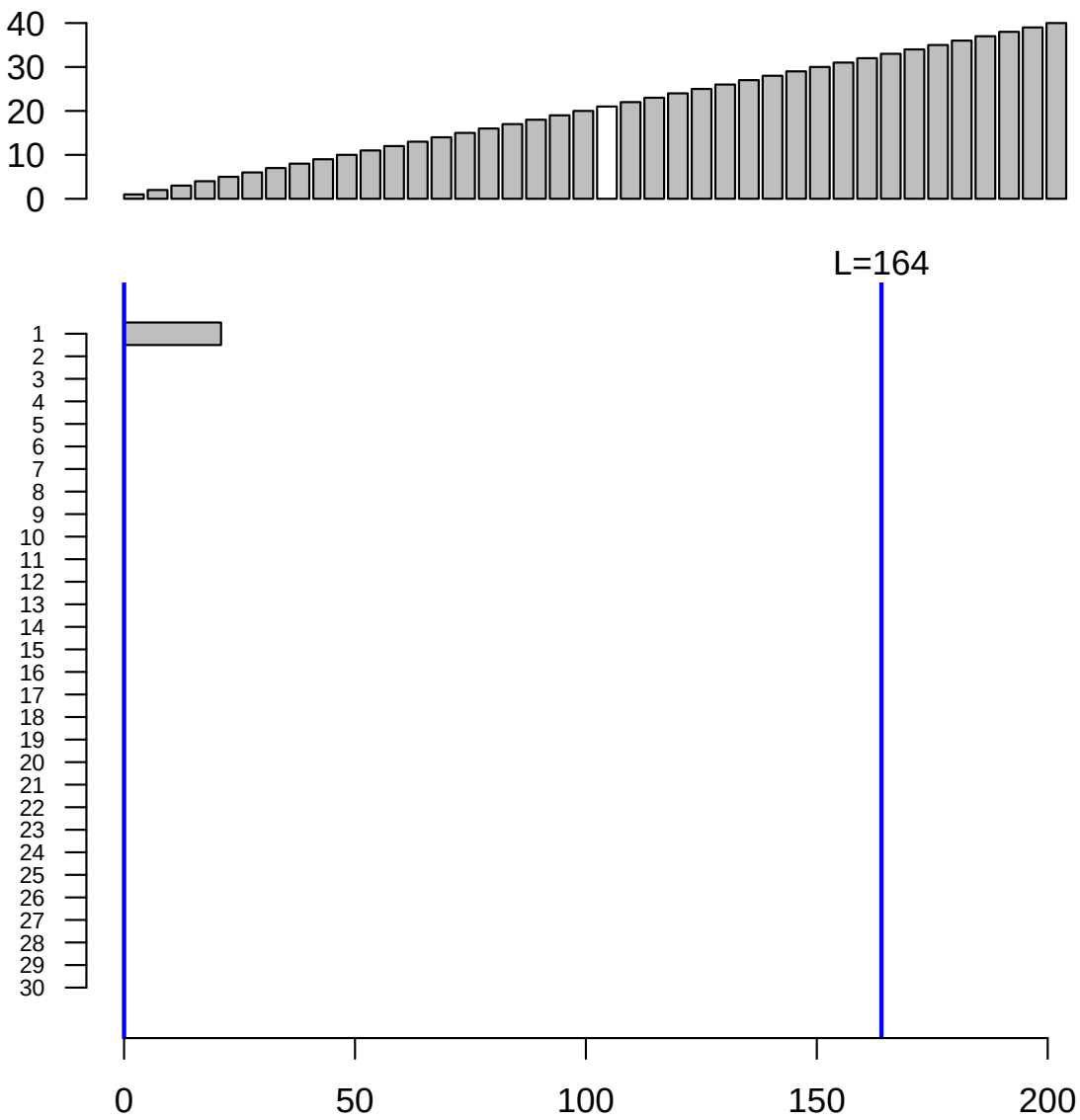
Example 6

How the budget is spent ? (SRSWOR, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



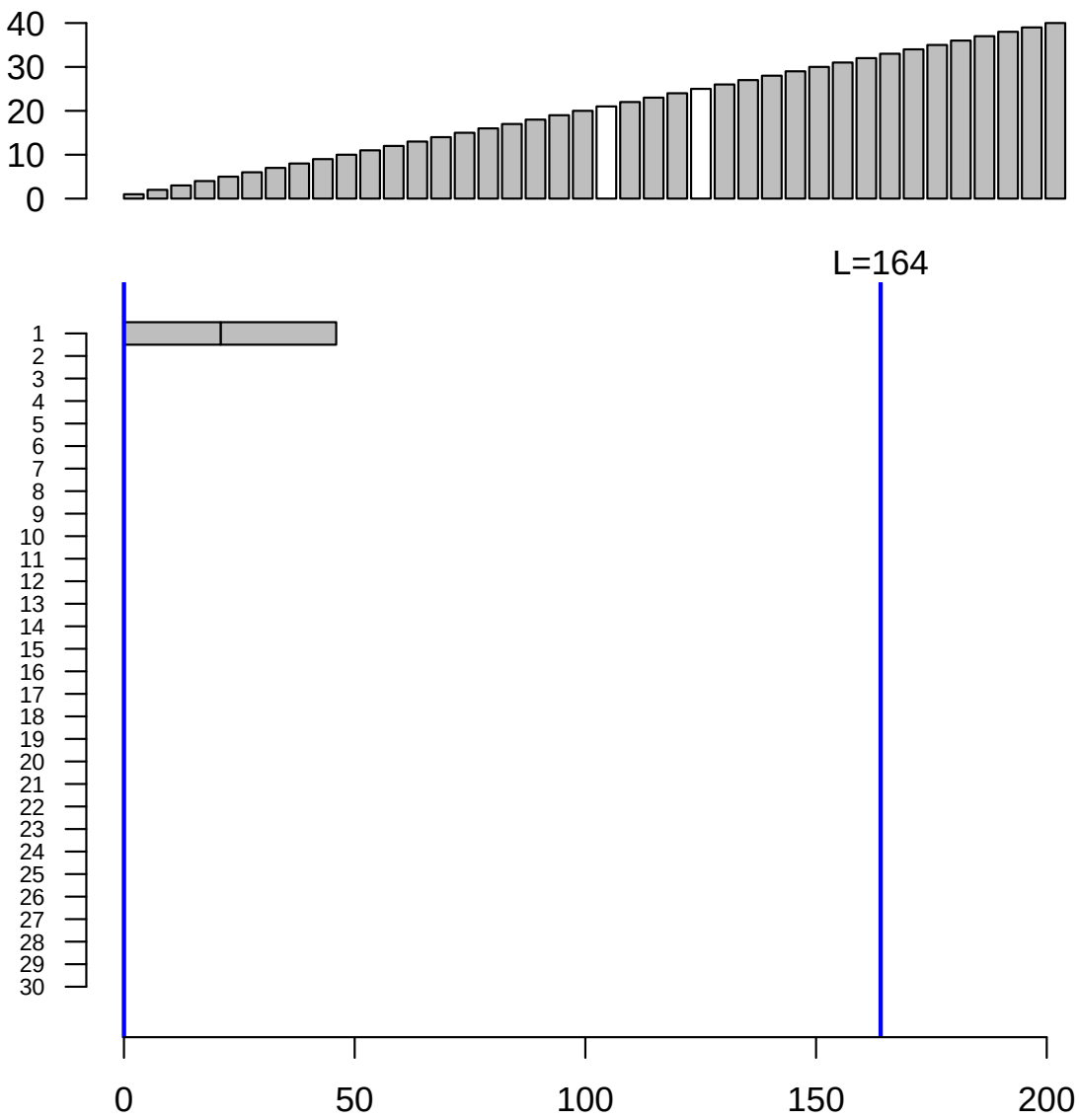
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



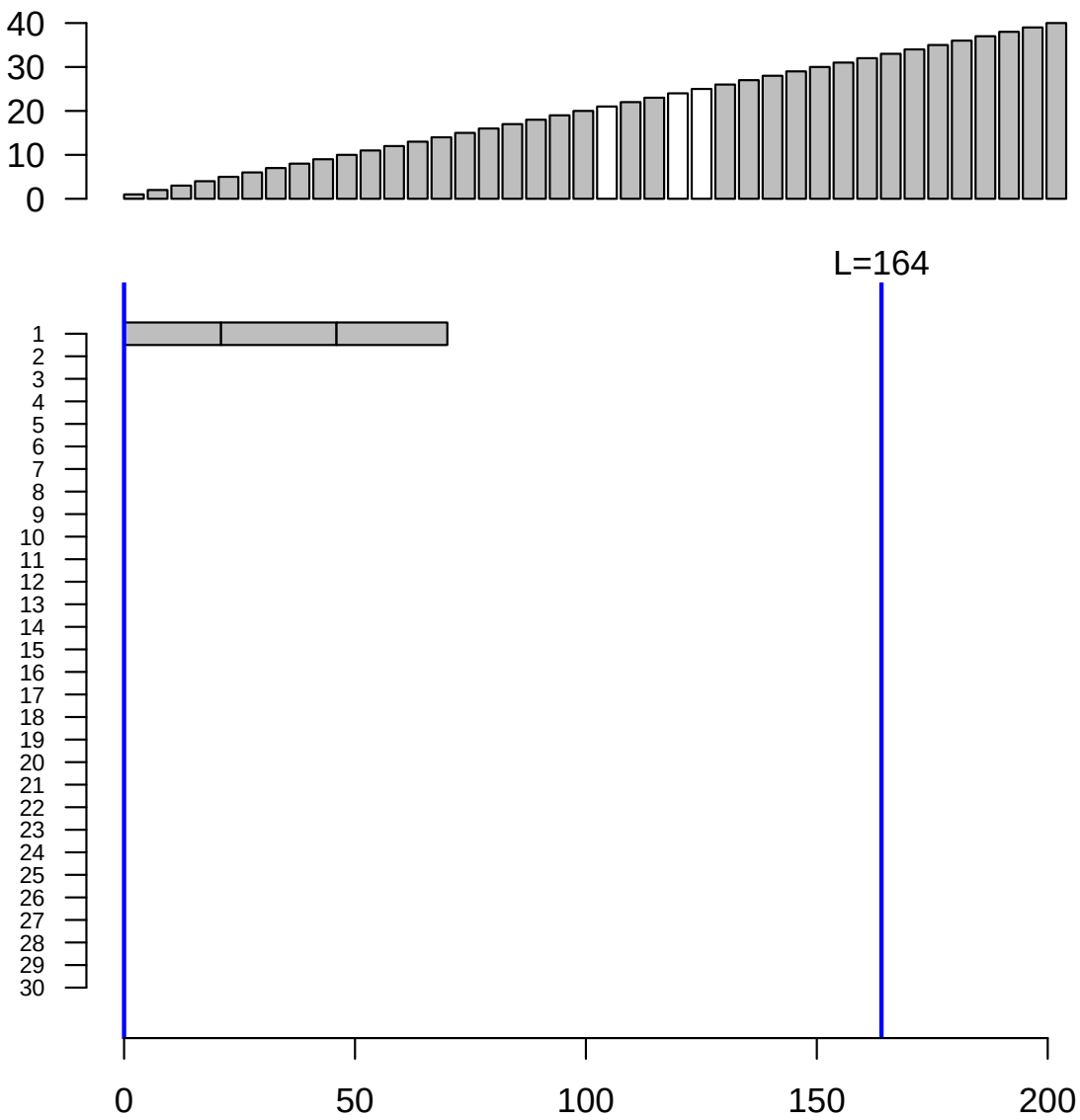
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



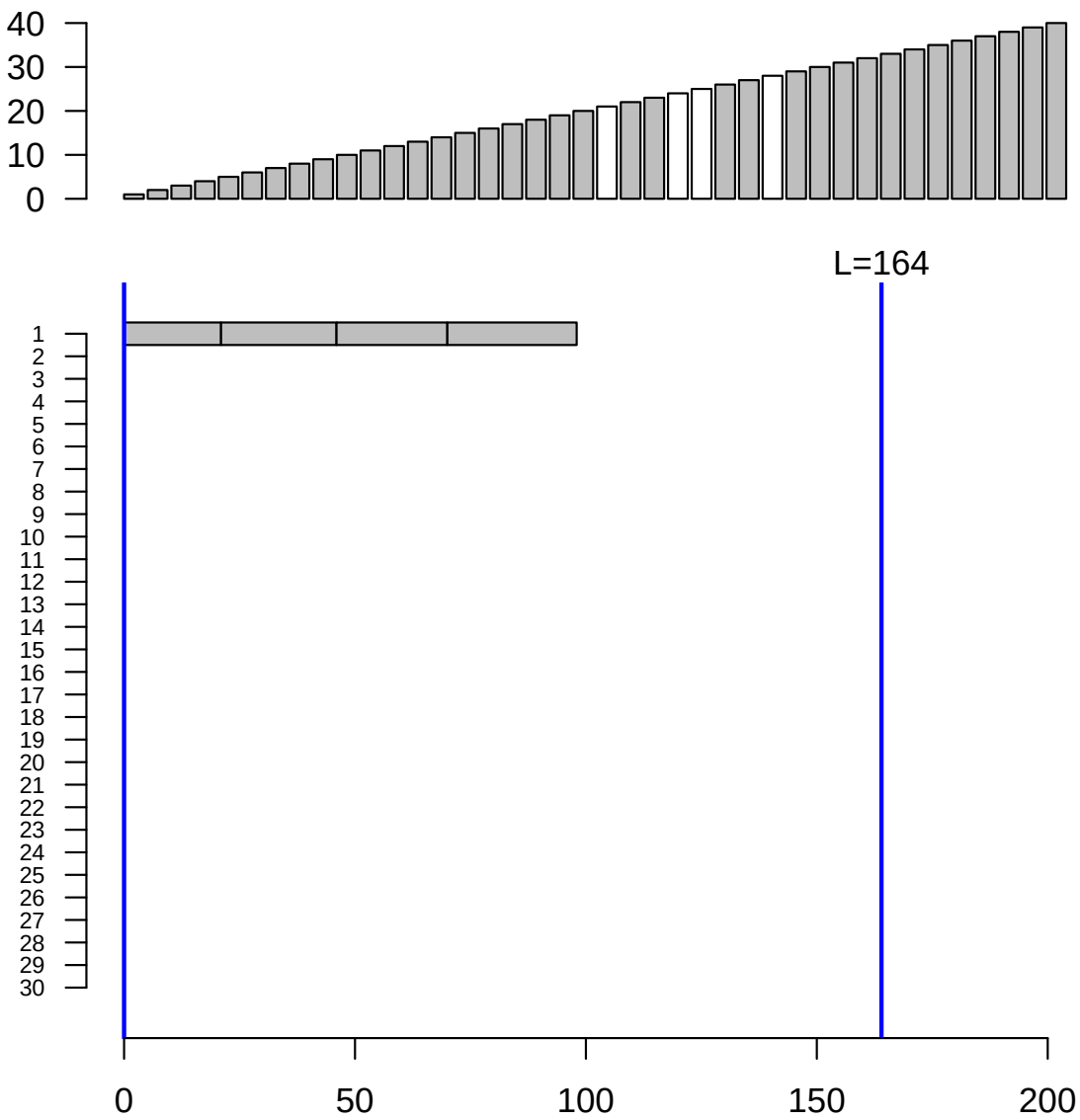
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



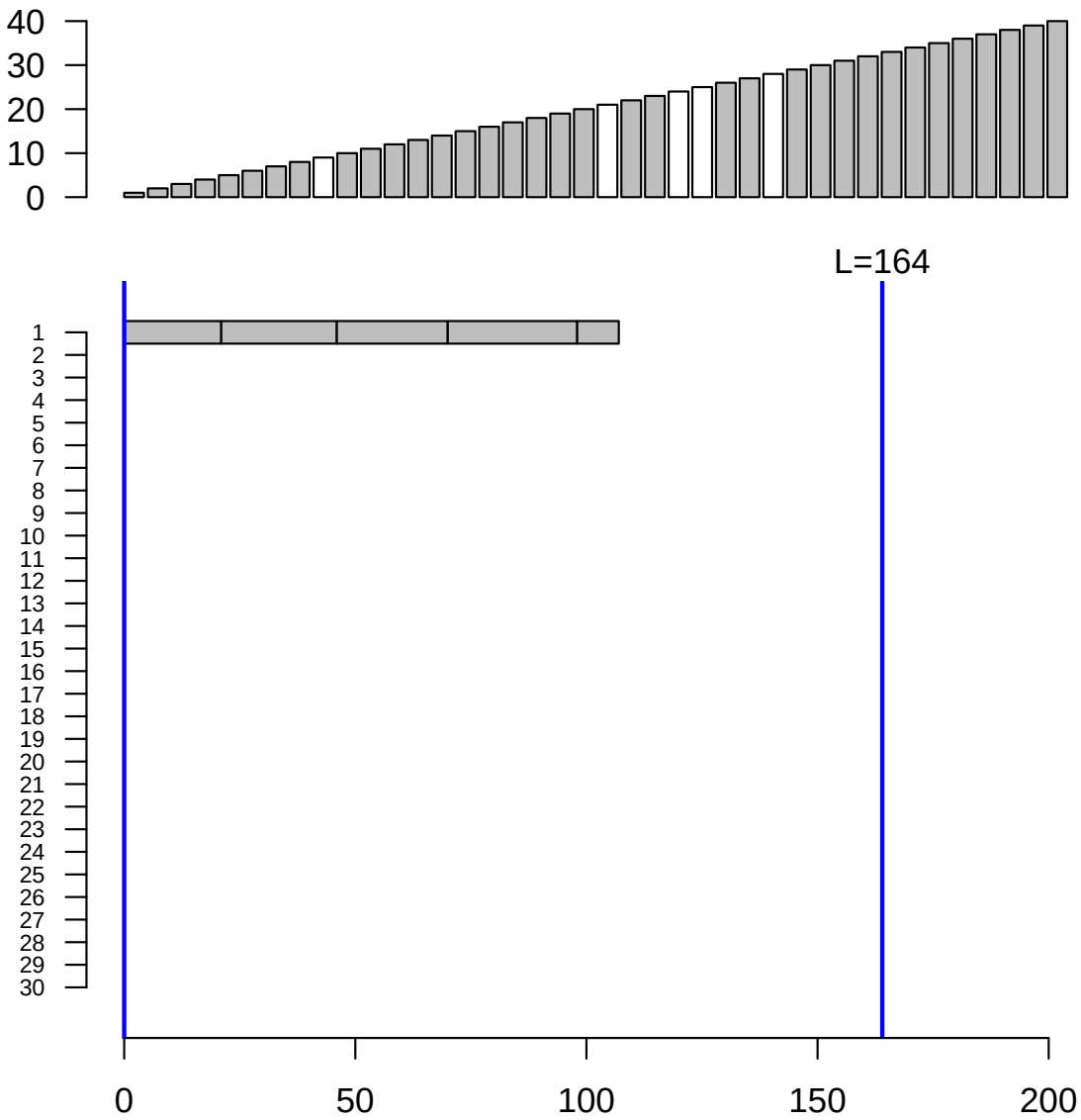
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



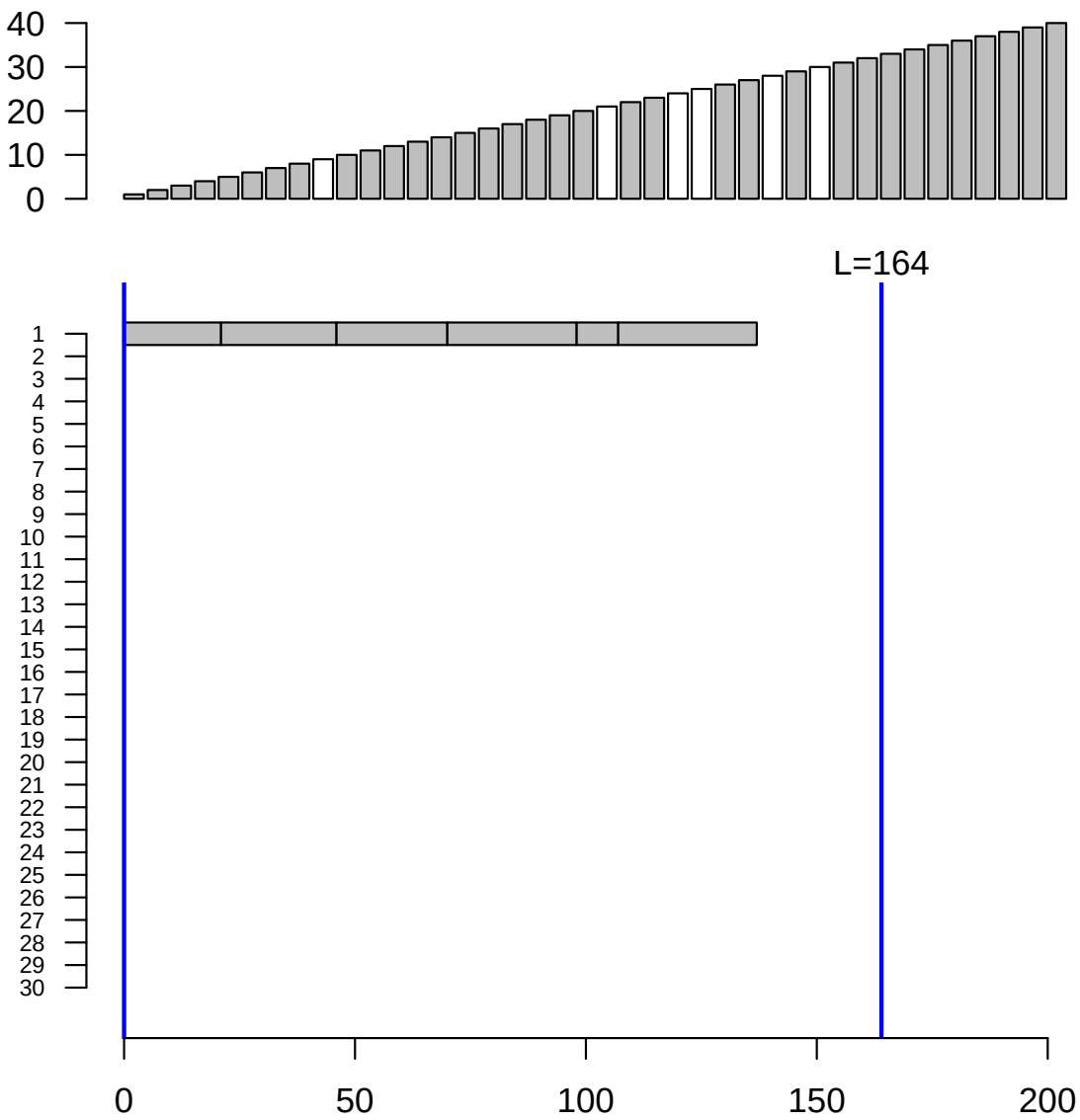
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



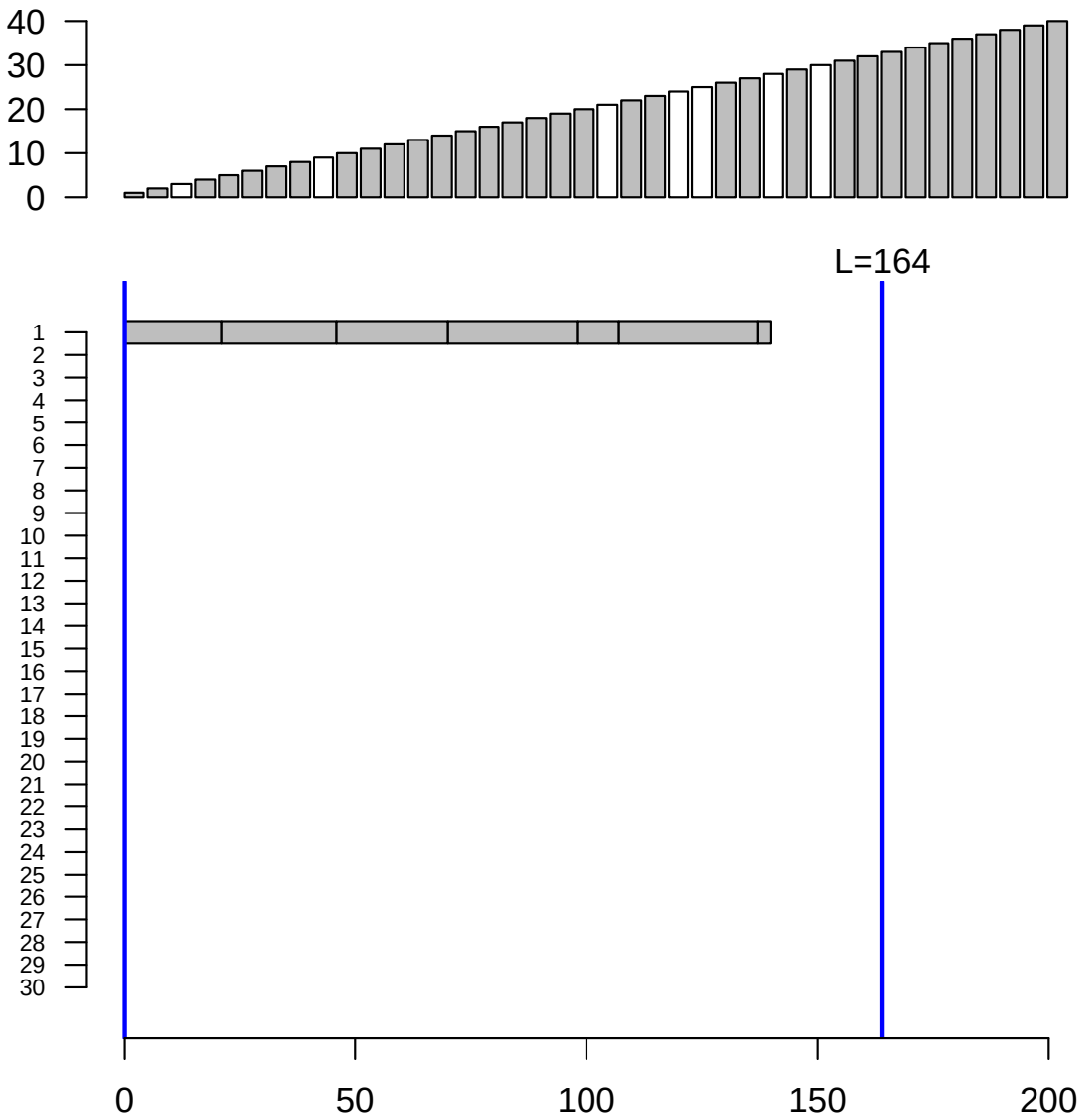
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



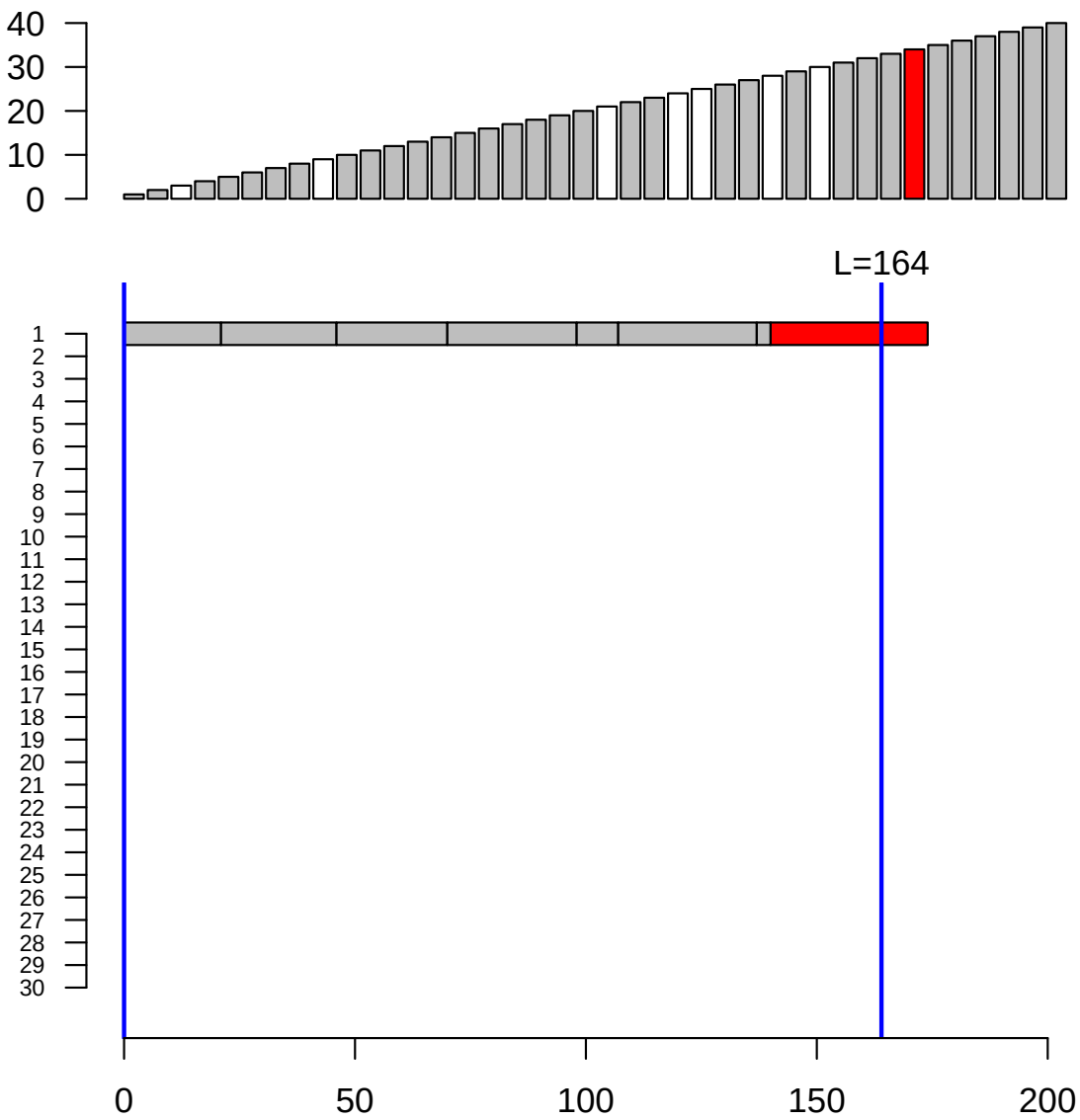
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



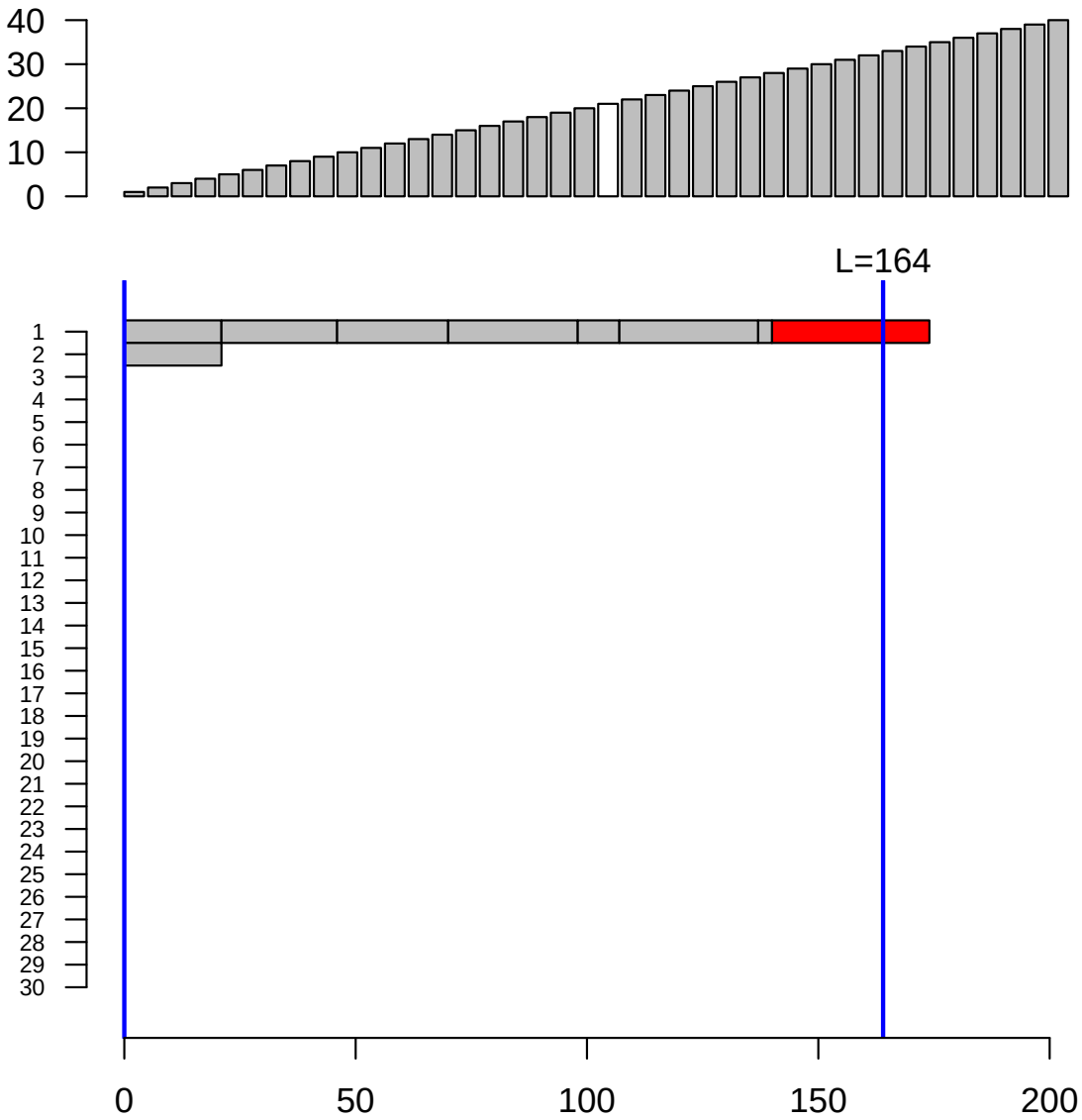
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



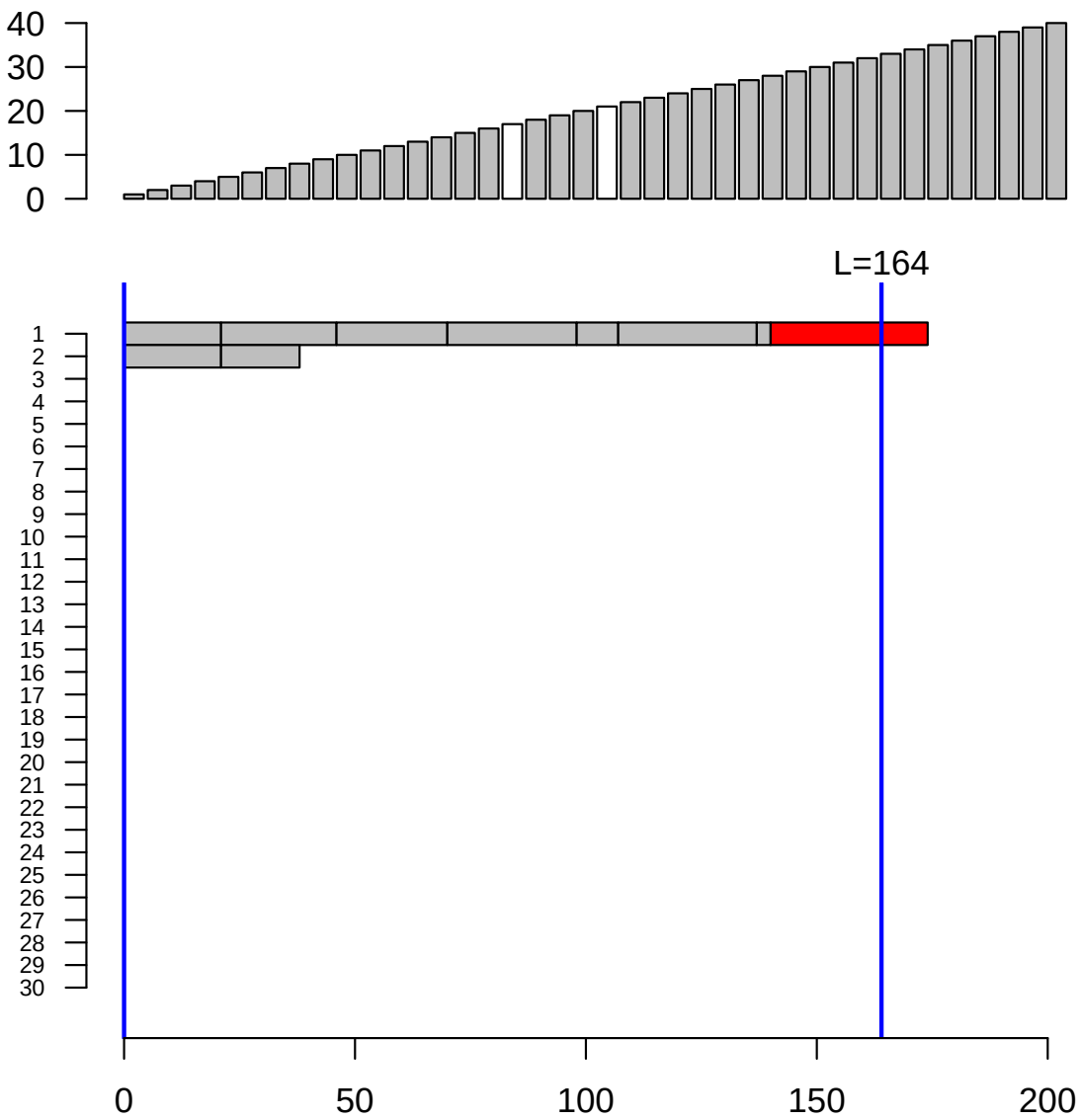
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



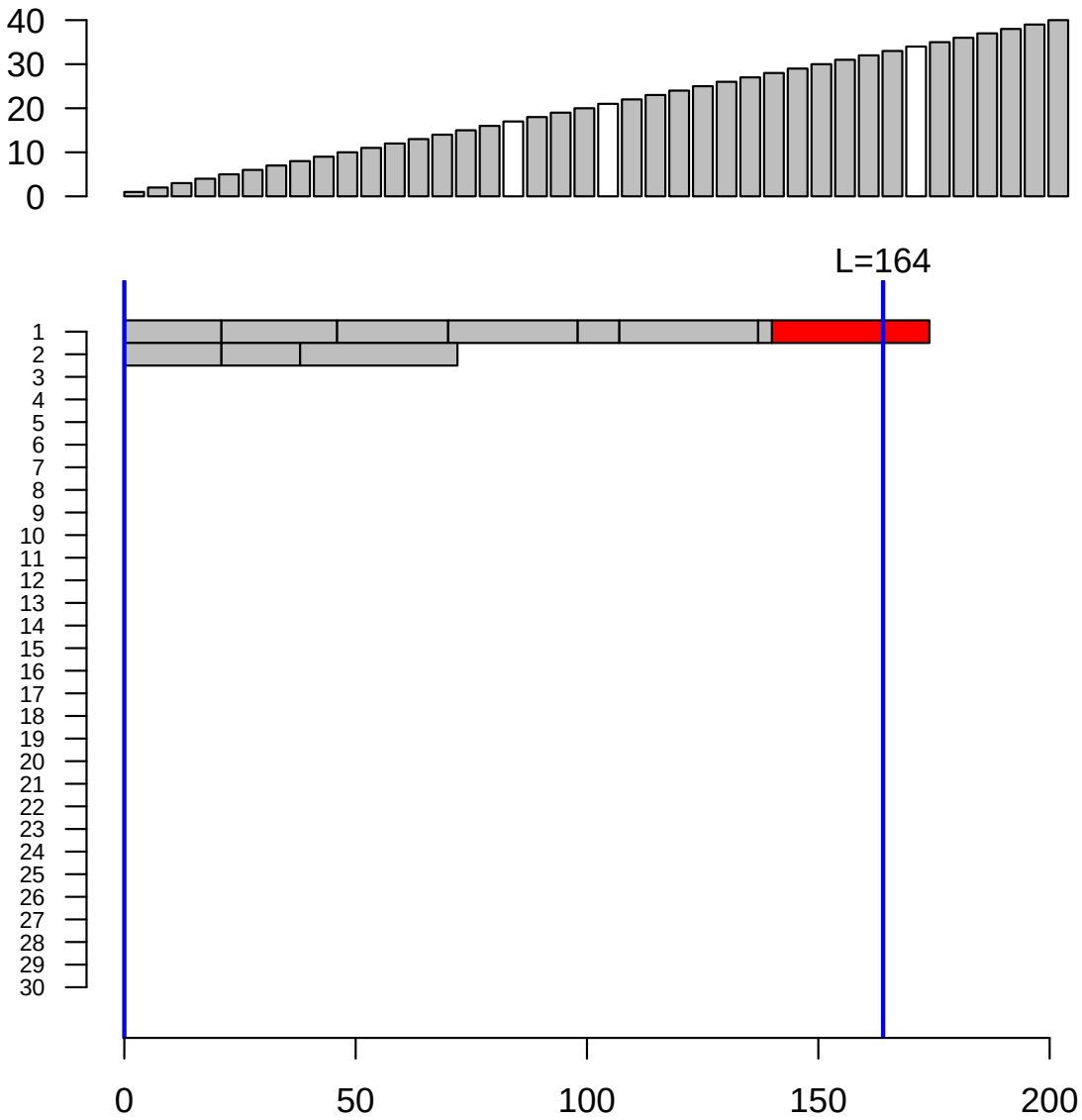
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



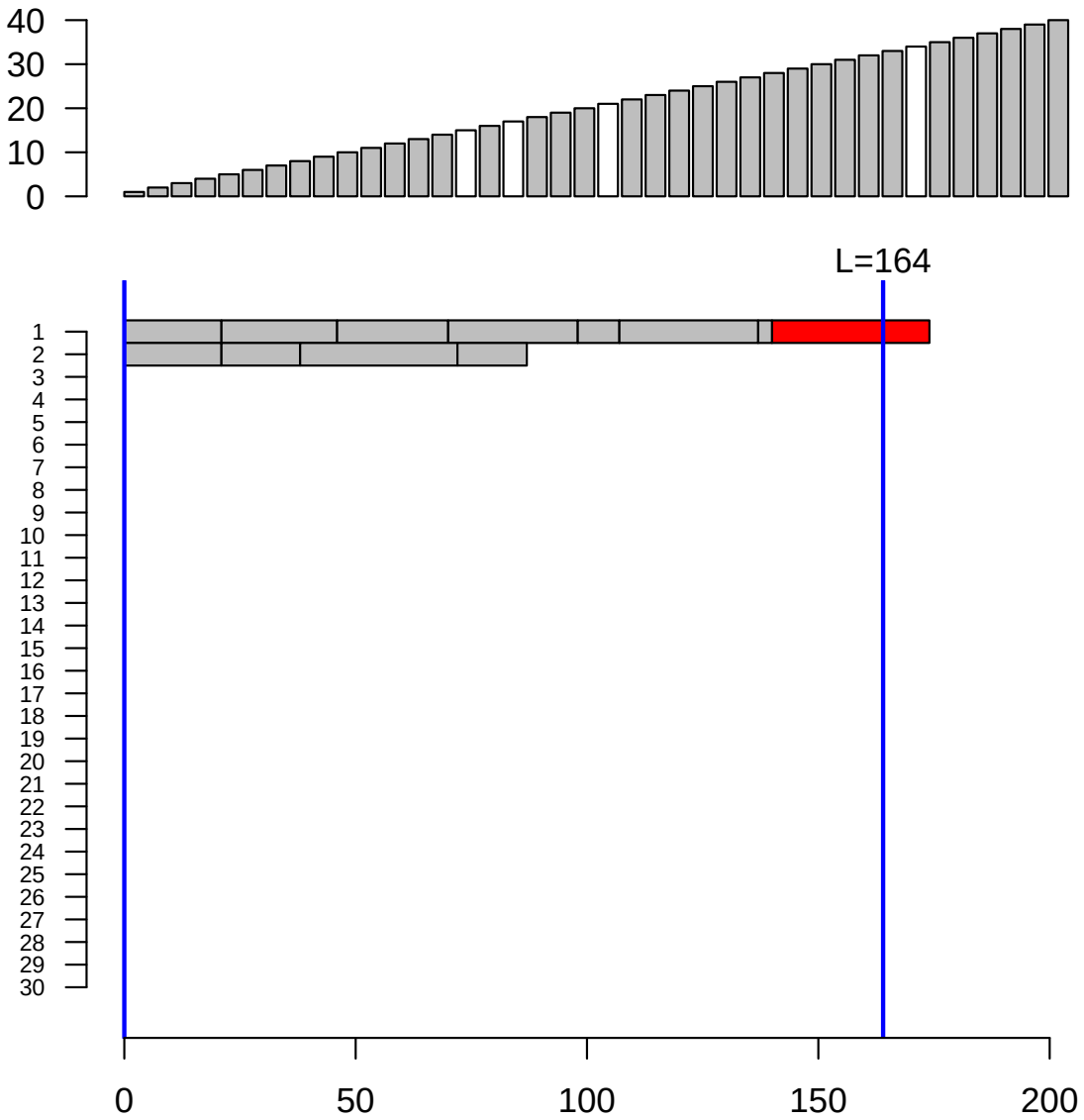
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



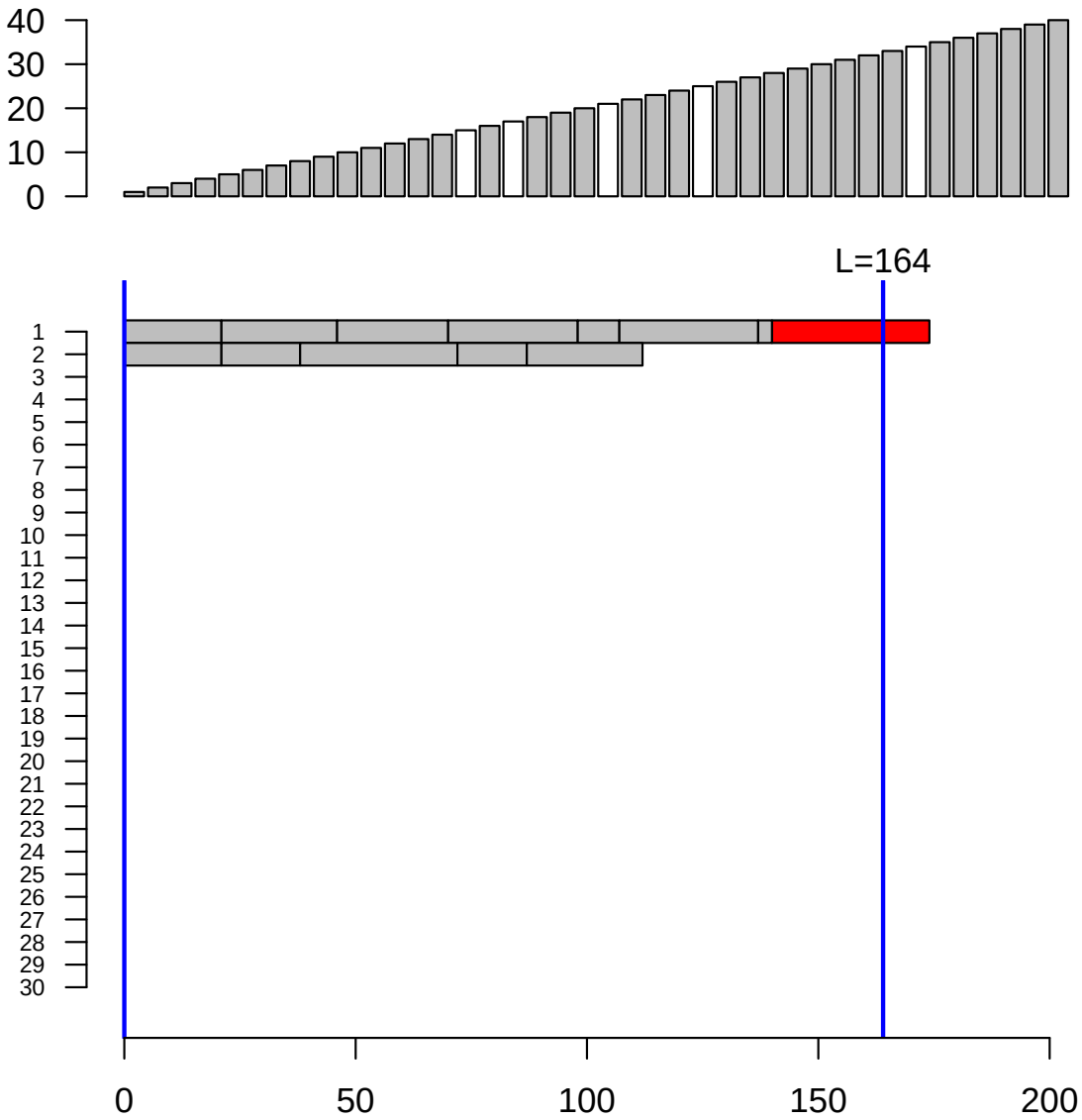
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



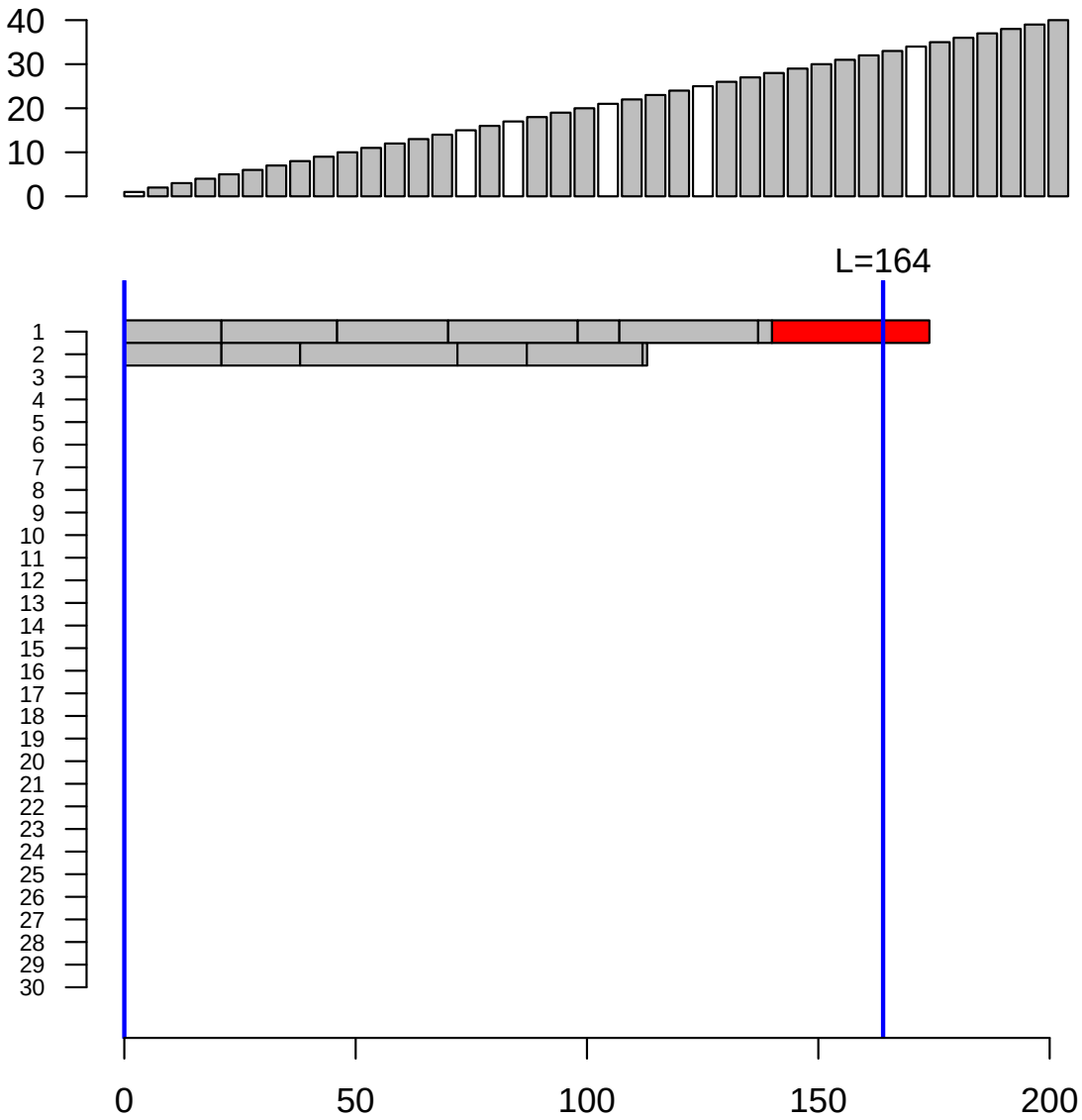
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



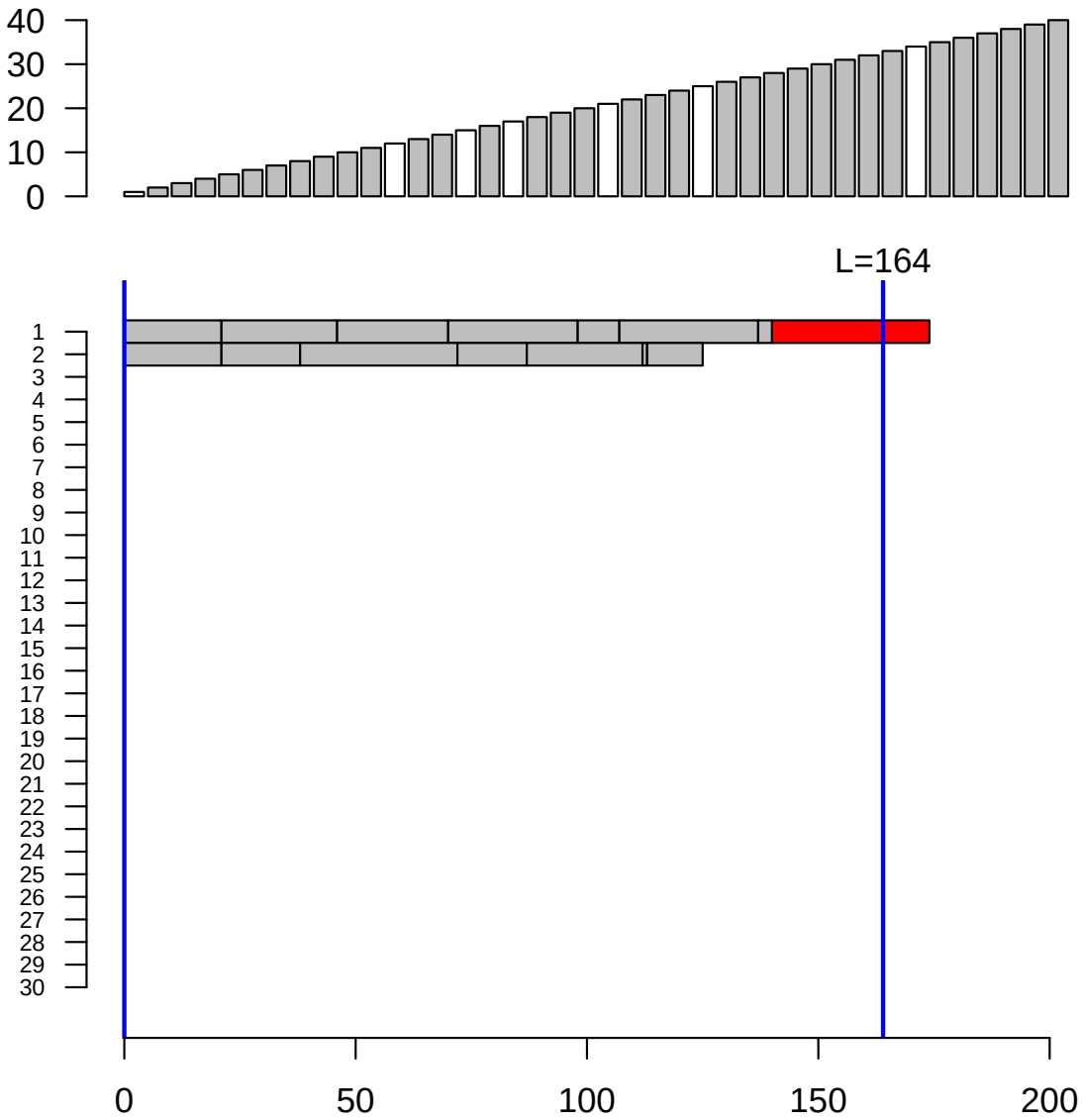
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



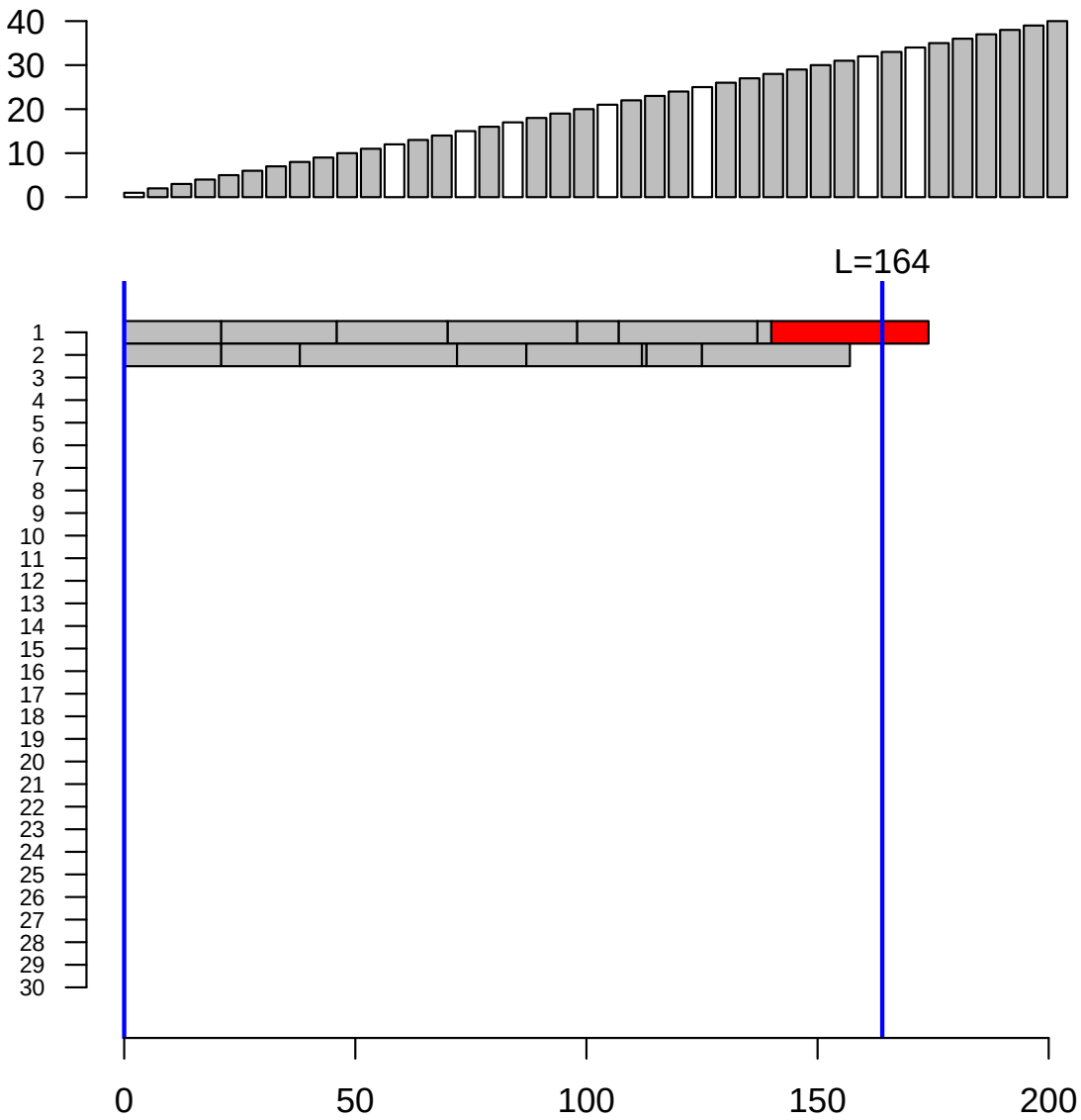
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



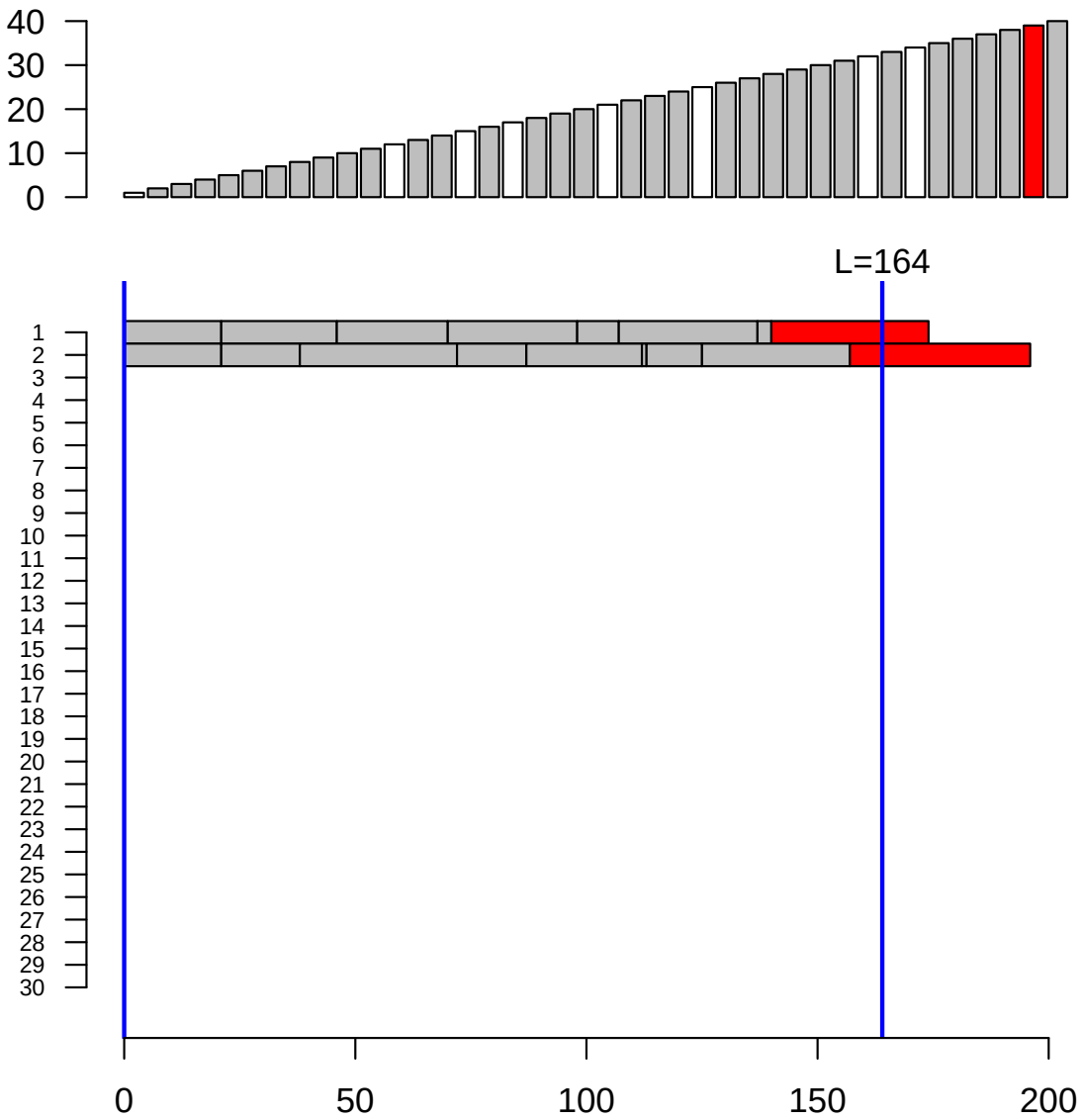
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



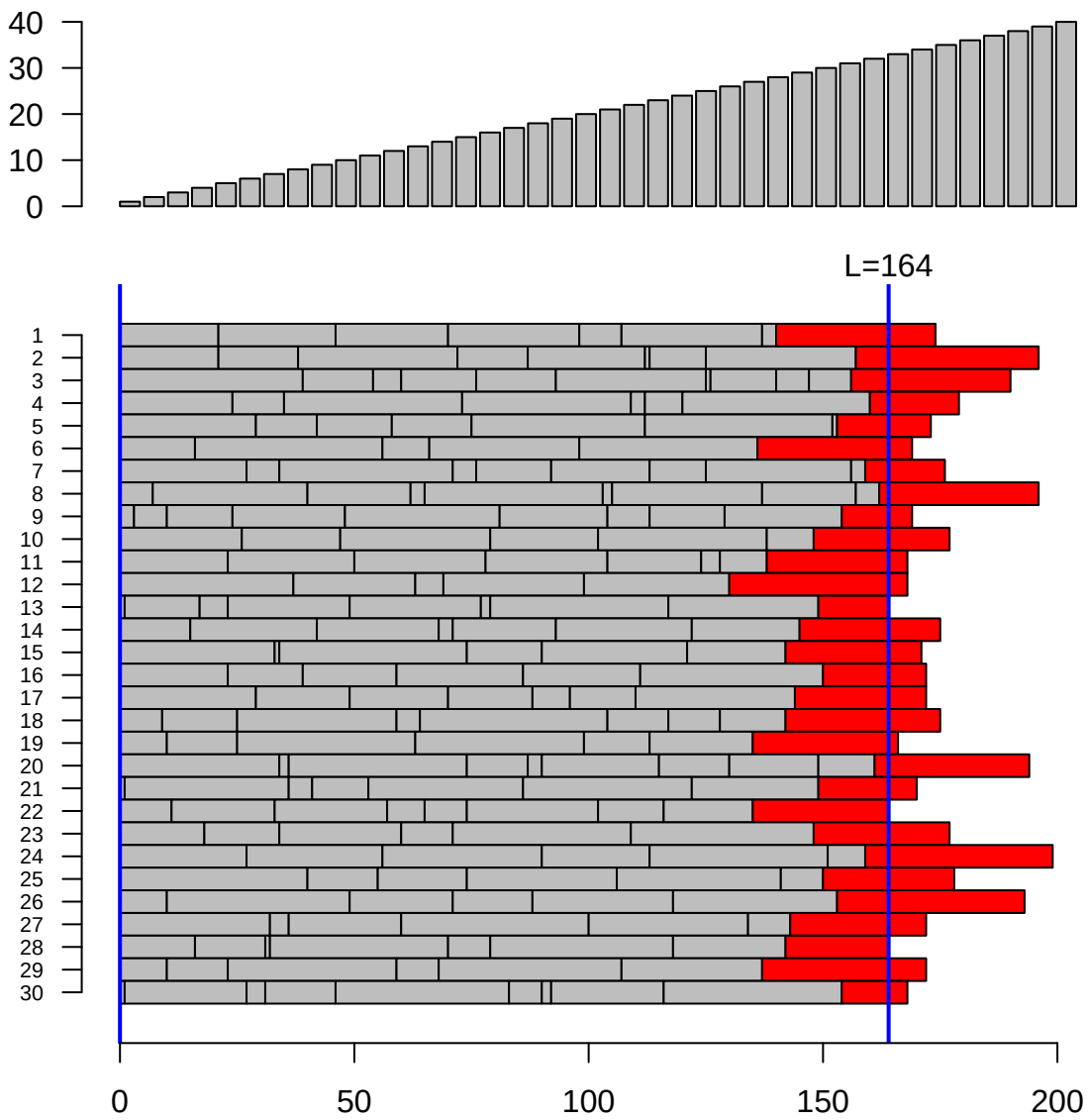
Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



Example 6

How the budget is spent ? (Pathak, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



A greedy sampling scheme (Gamrot 2015)

For any set $Q \subseteq U$ denote:

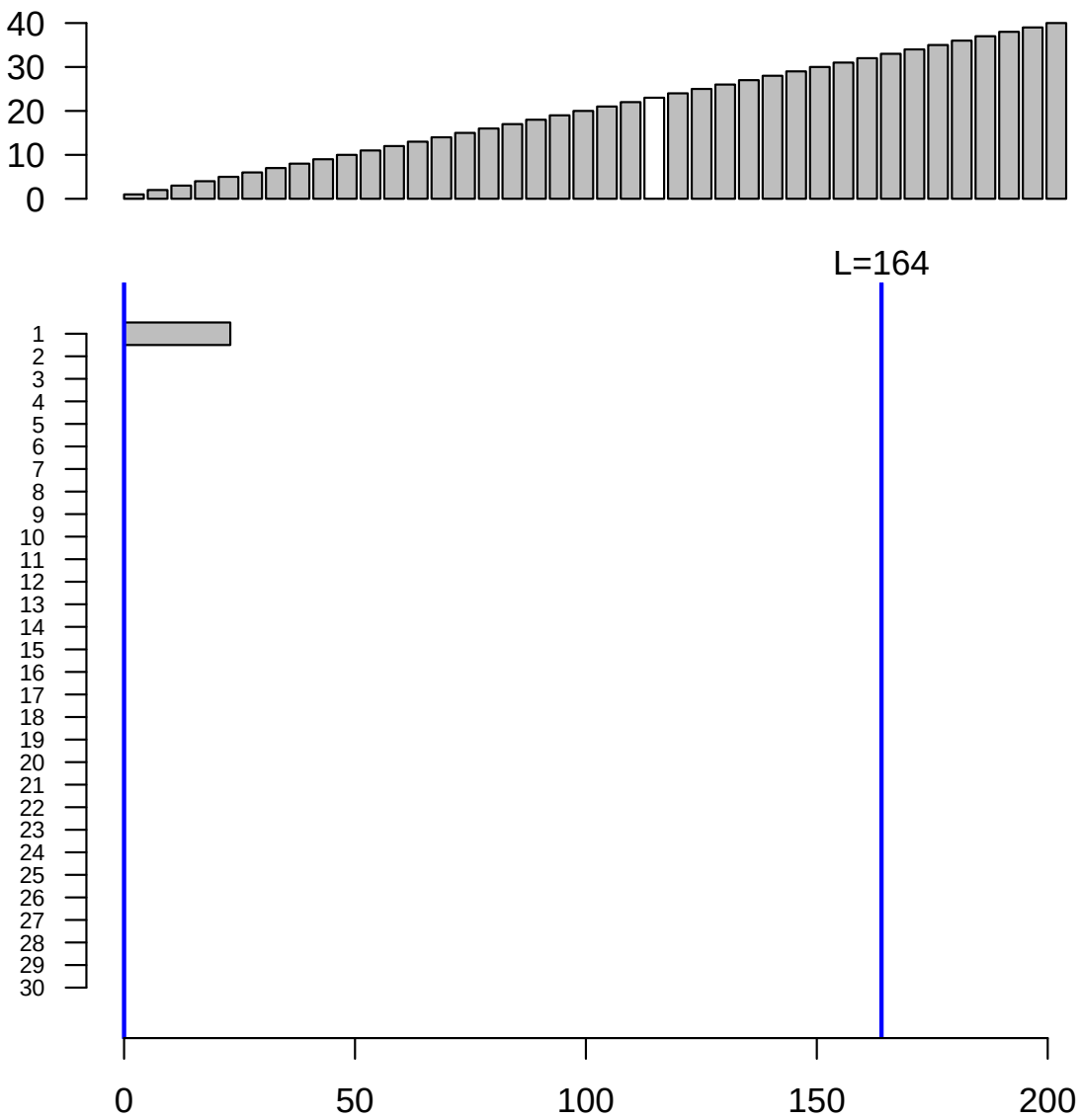
$$U(Q) = \left\{ i \in U - Q : c_i \leq L - \sum_{i \in Q} c_i \right\} \tag{11}$$

The scheme works as follows:

1. Let $s_0 = \emptyset$
2. For $k = 1, 2, \ldots$ do the following:
   - If $U(s_{k-1})$ is nonempty, then draw an element A_k from it with equal probabilities and let $s_k = s_{k-1} \cup \{A_k\}$
   - If in some K-th step $U(s_k)$ is empty then go to step 3
3. The set s_K is the sample

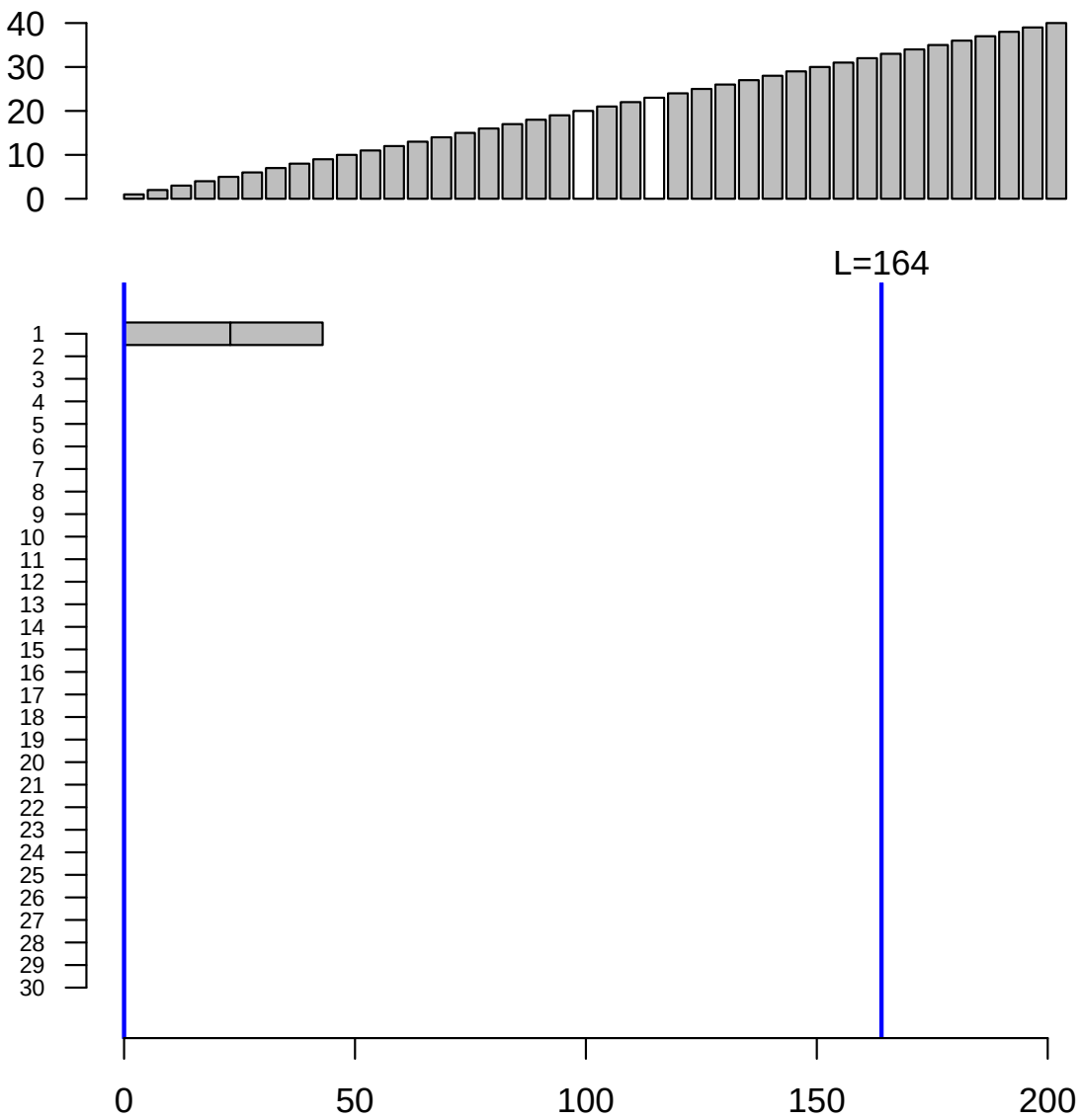
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



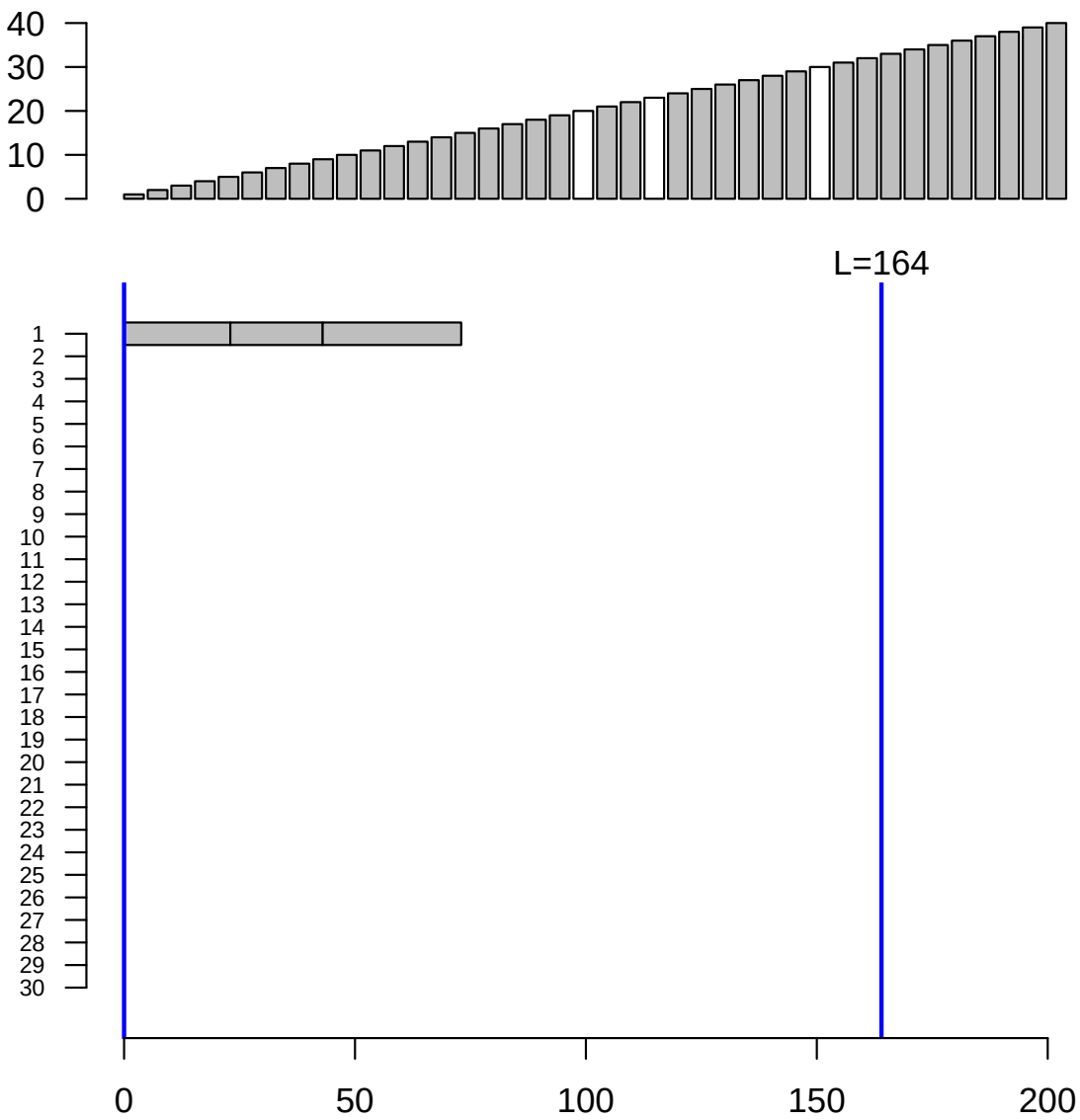
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



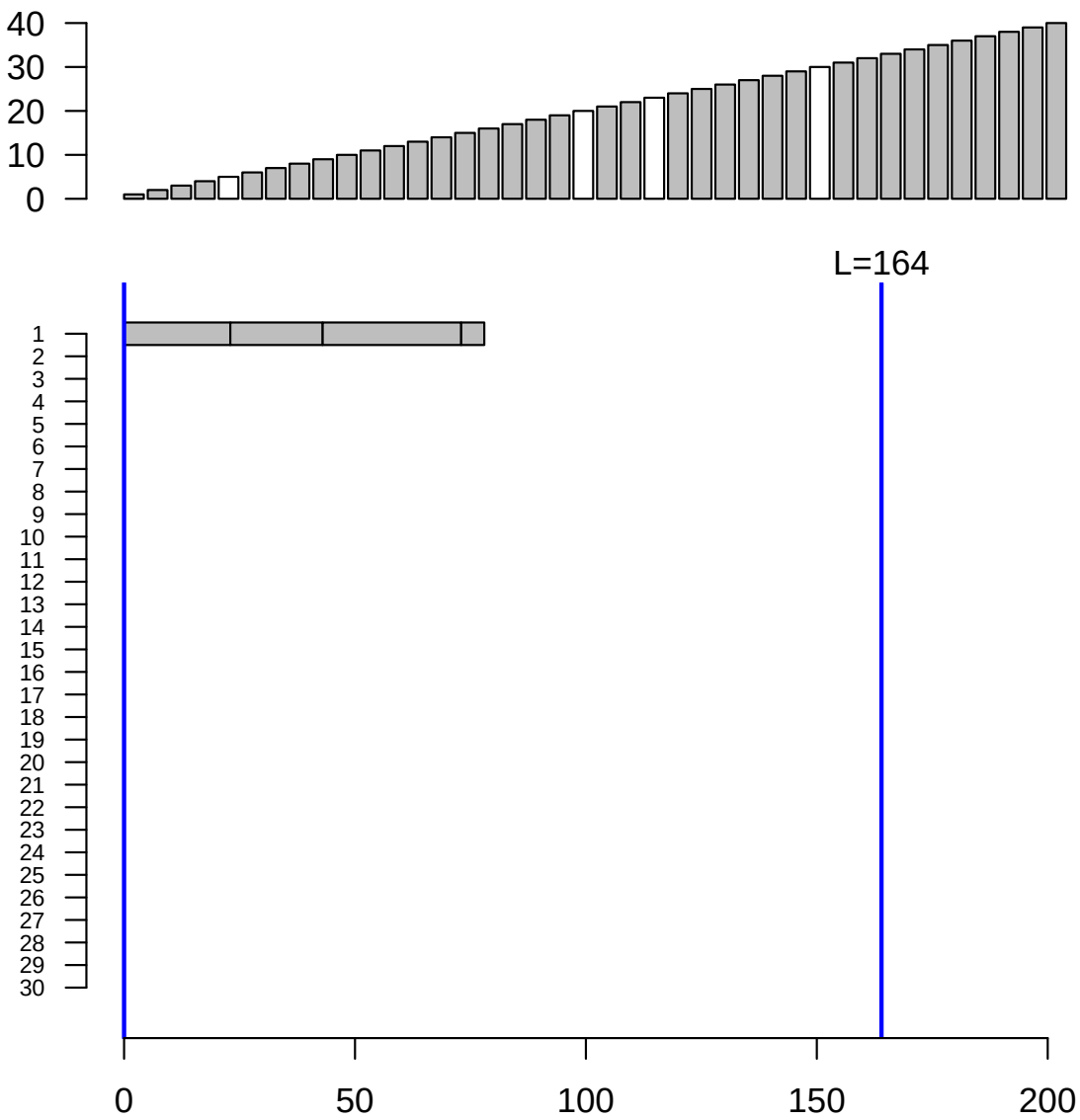
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



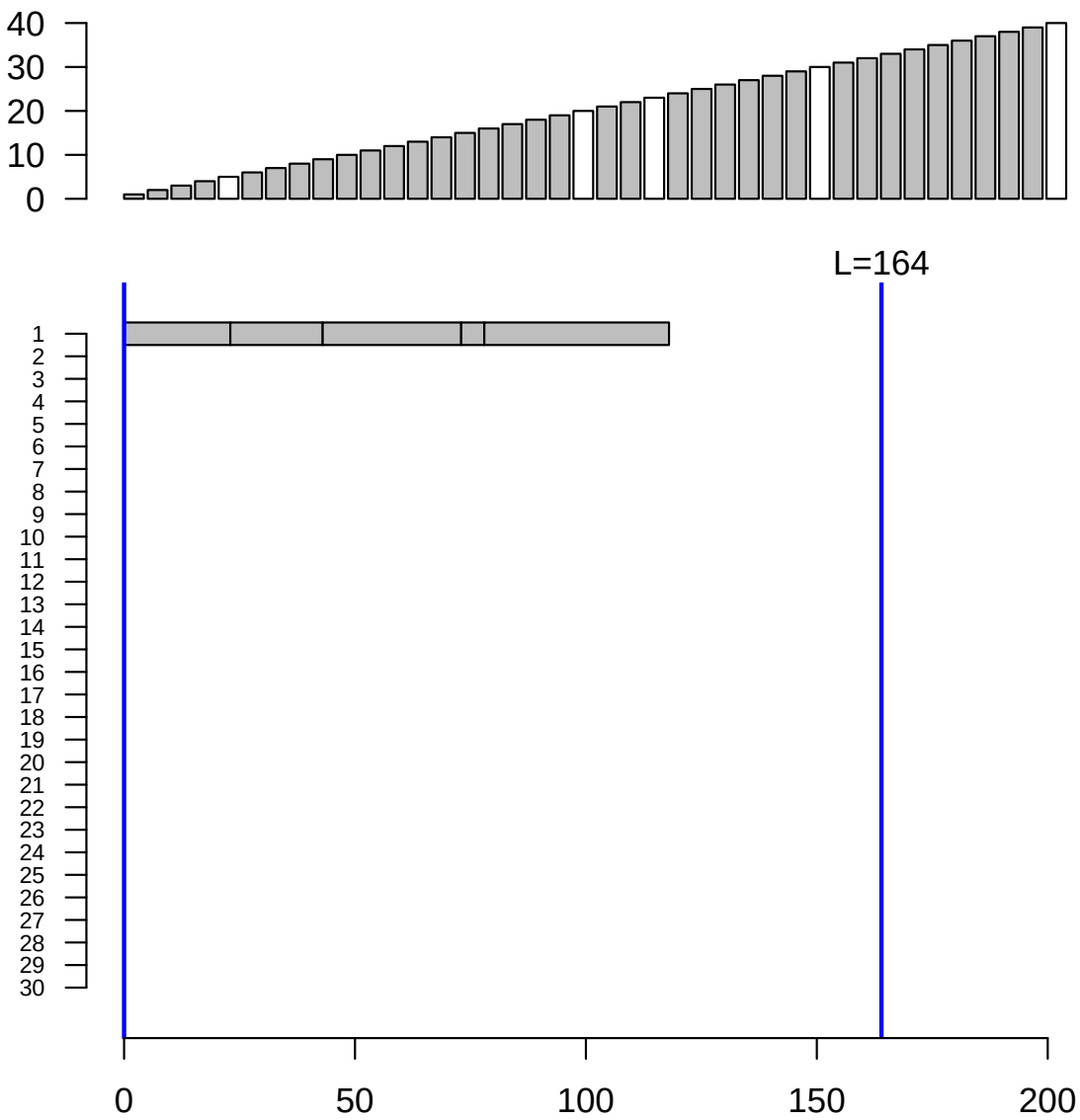
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



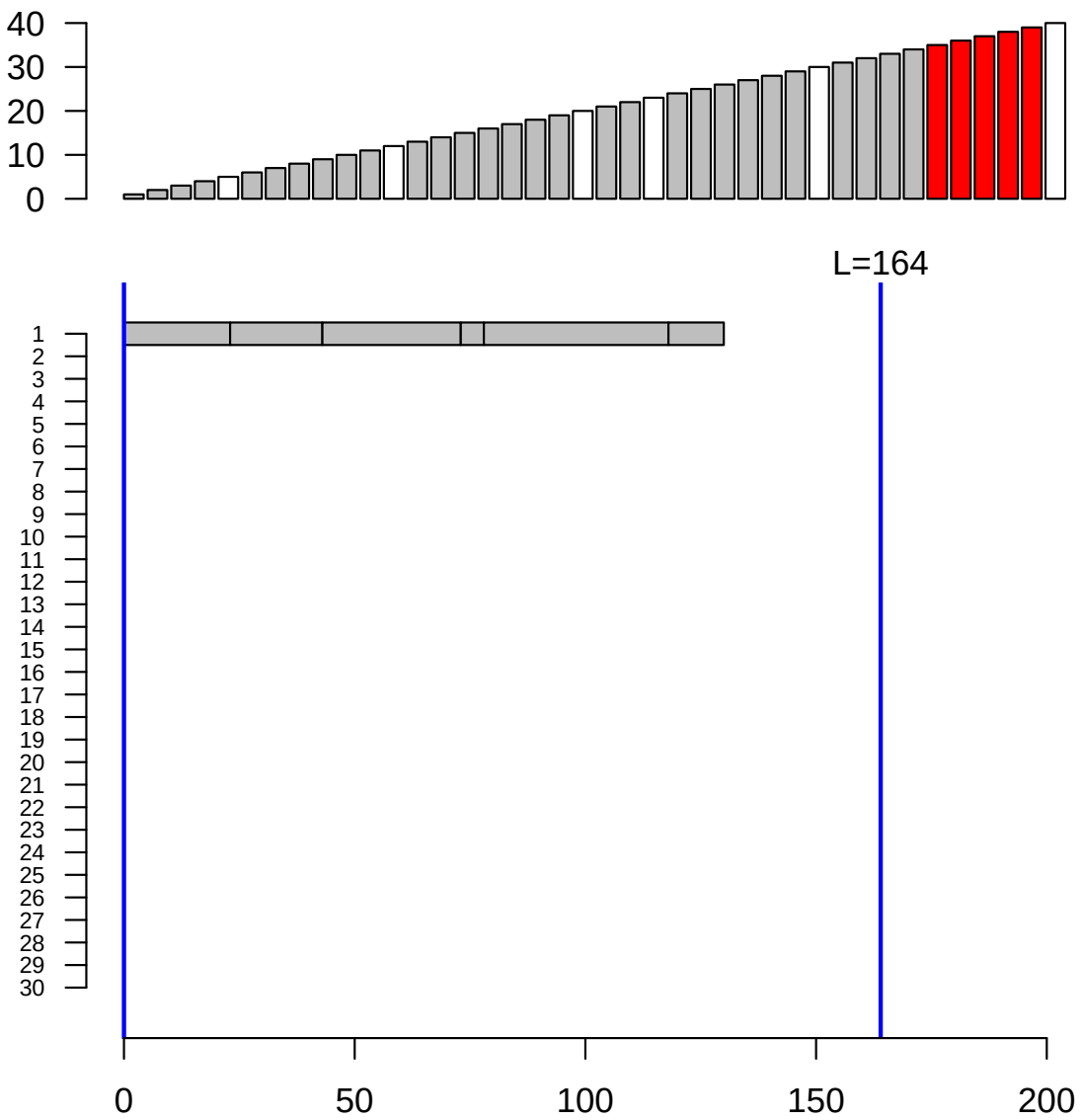
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



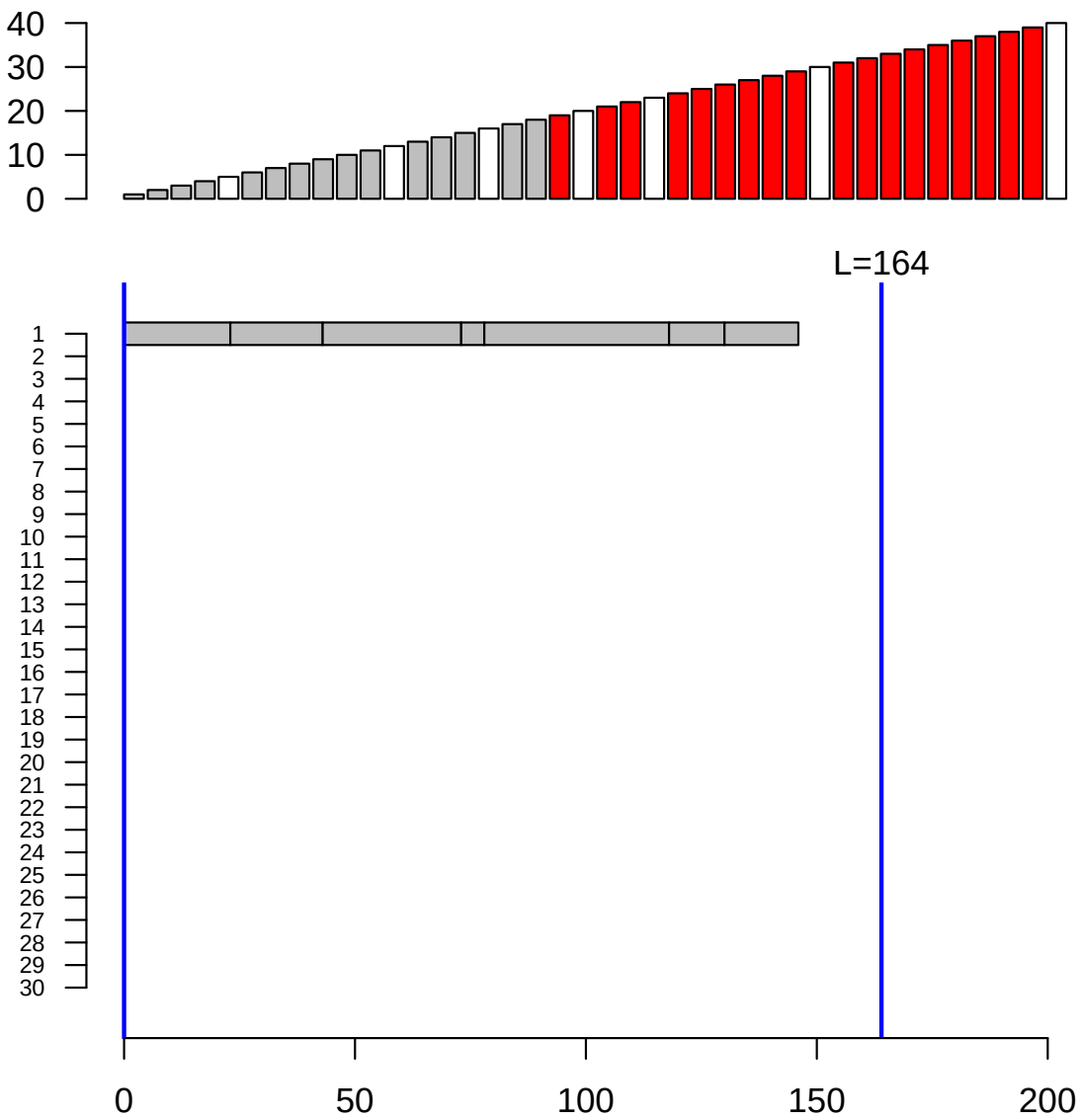
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



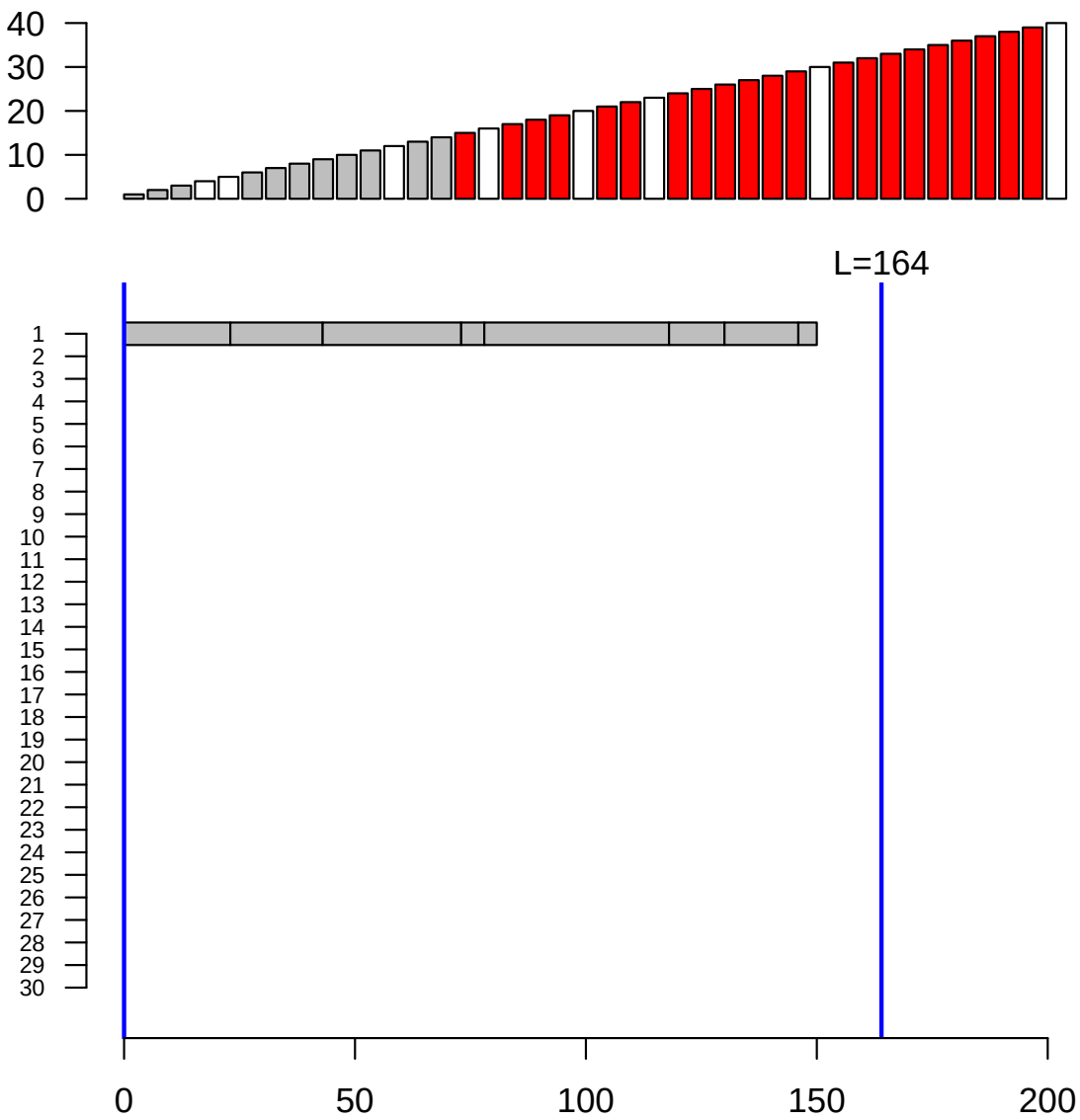
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



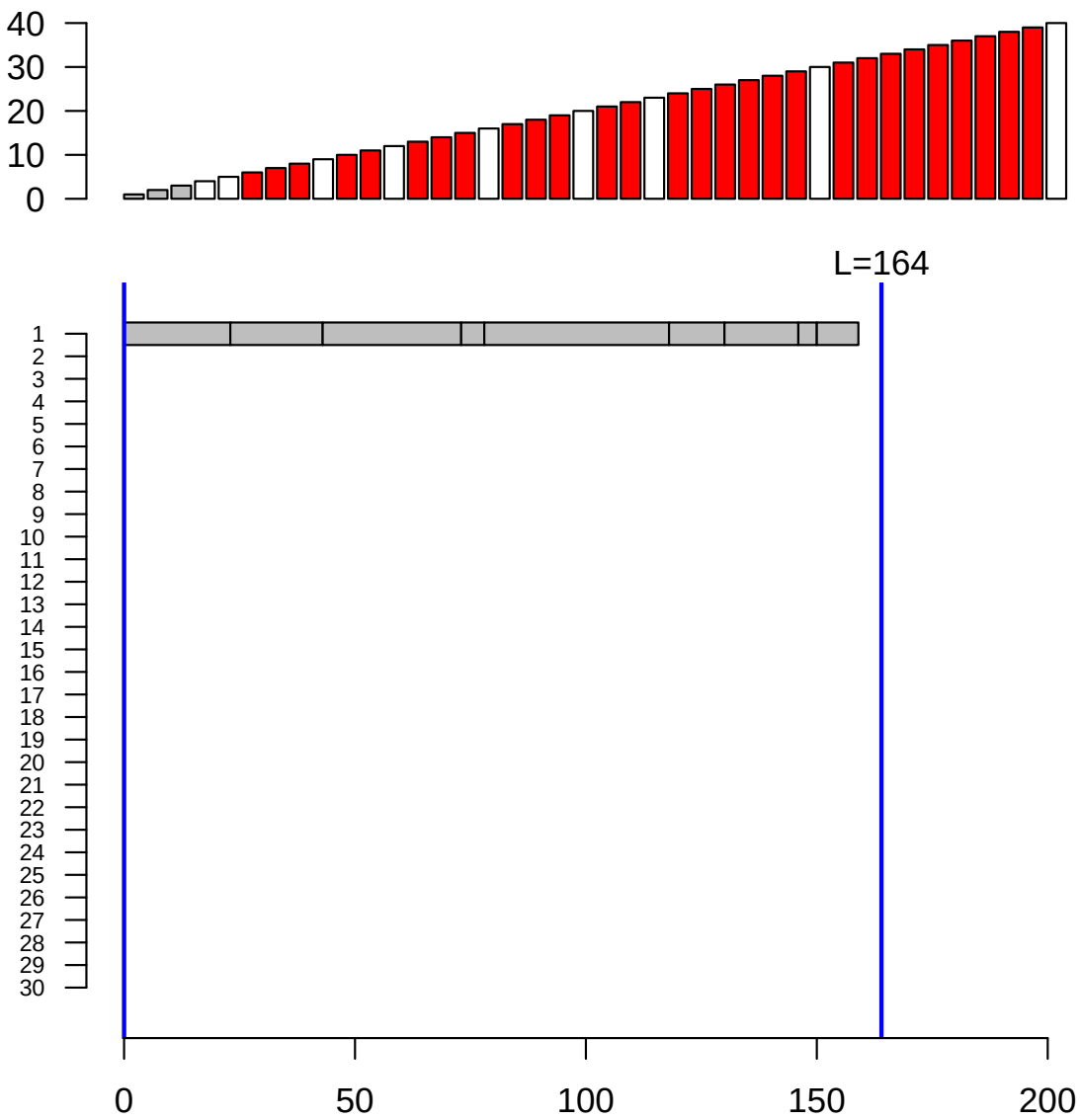
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



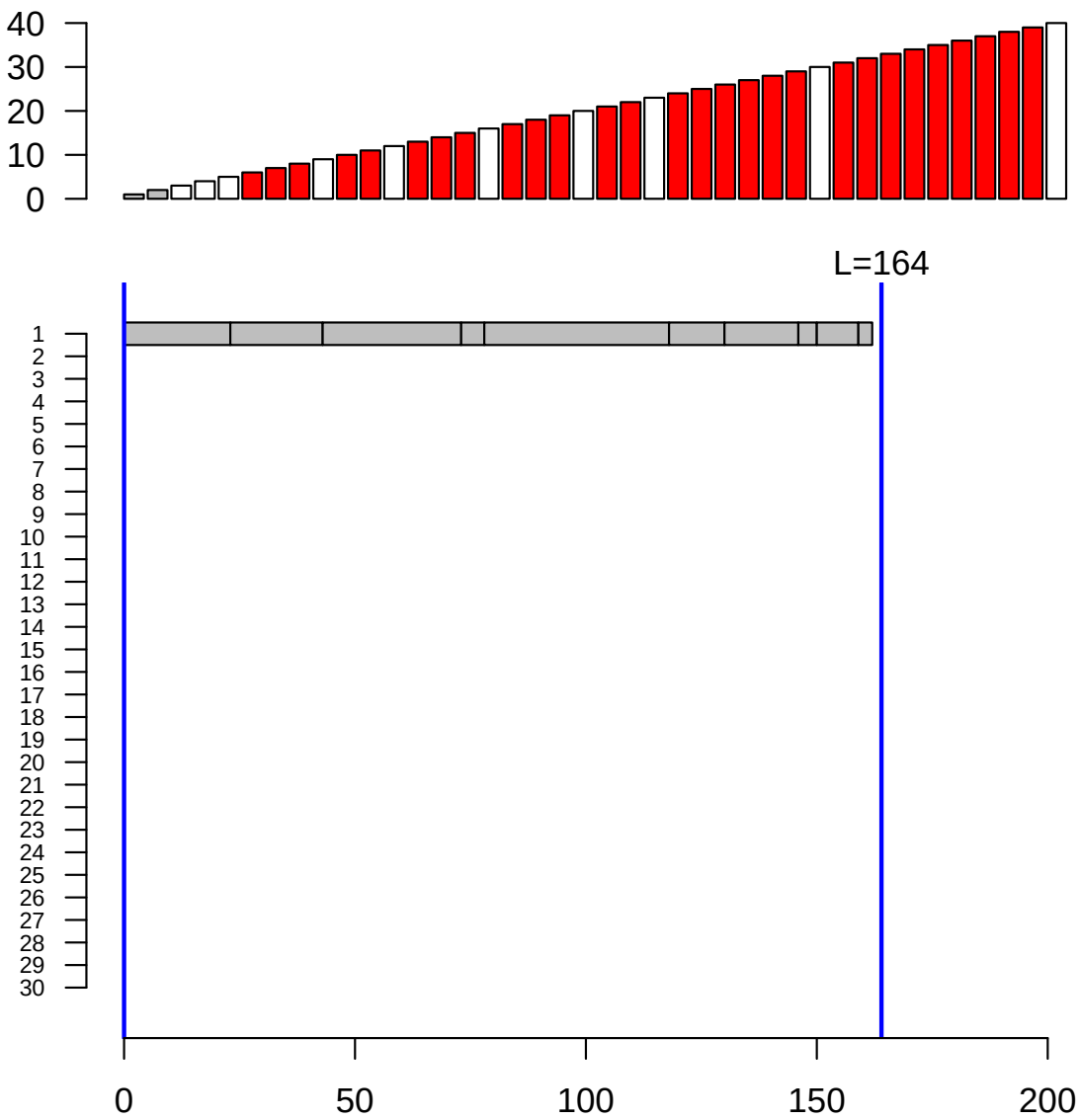
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



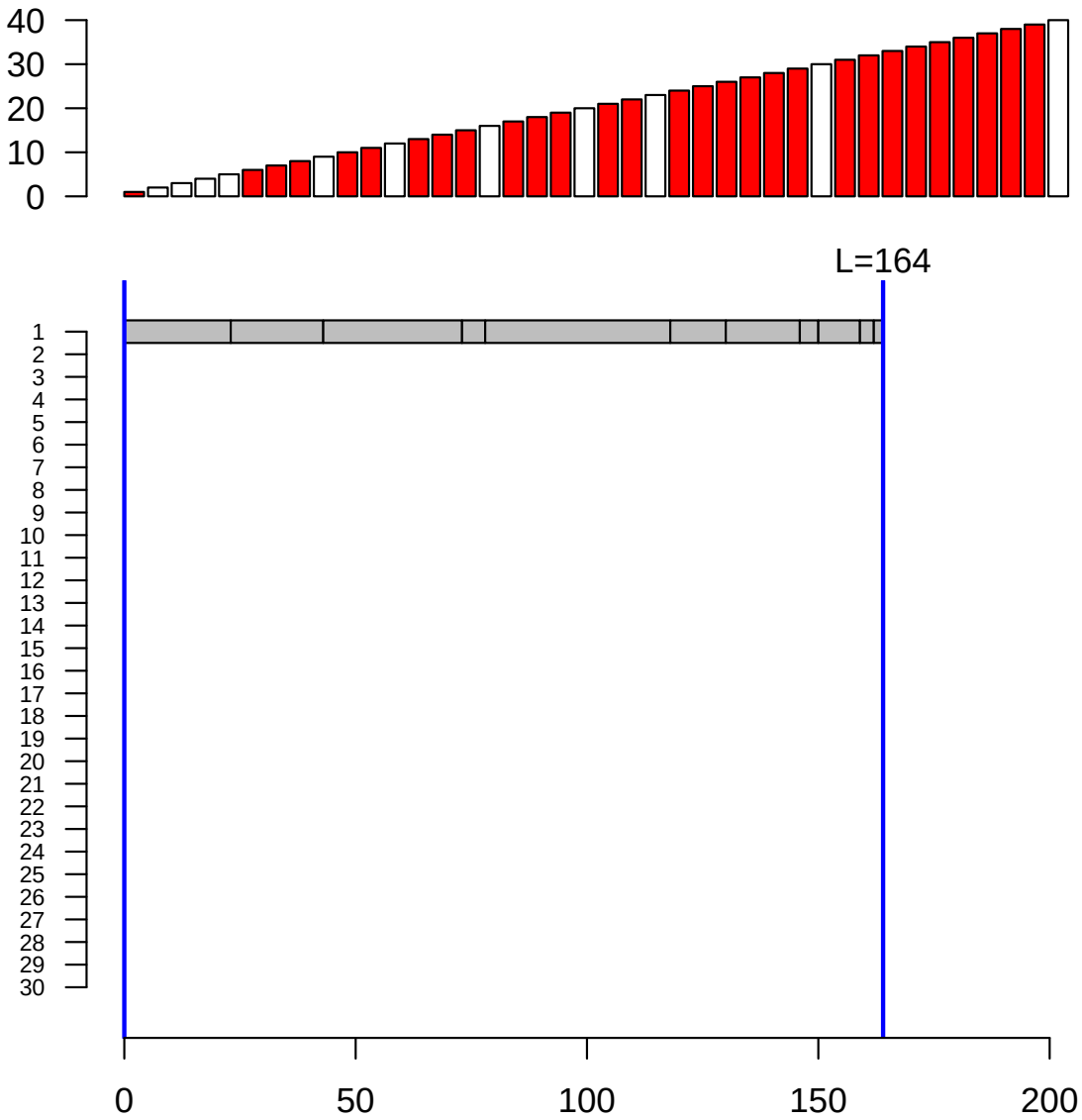
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



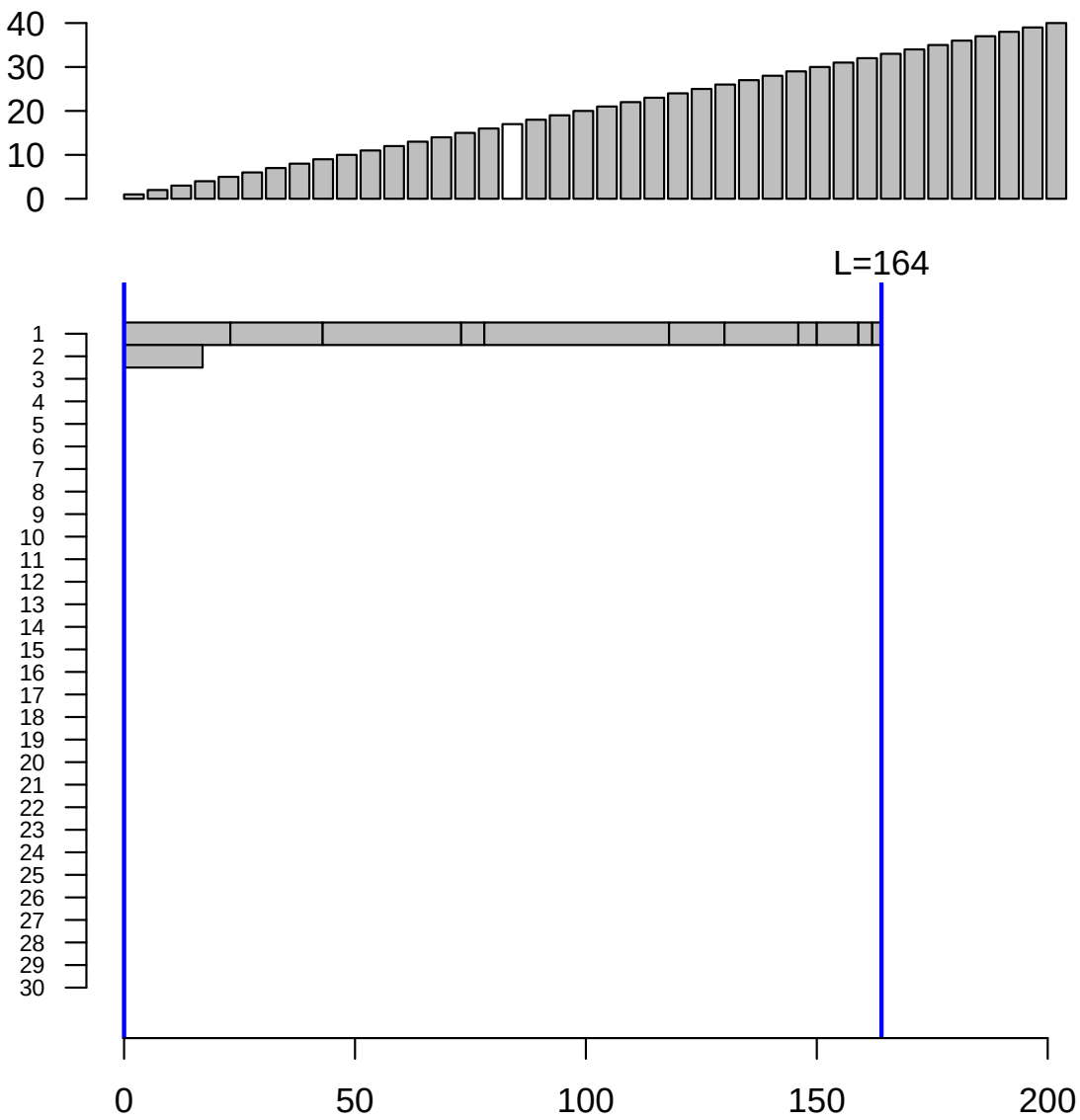
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



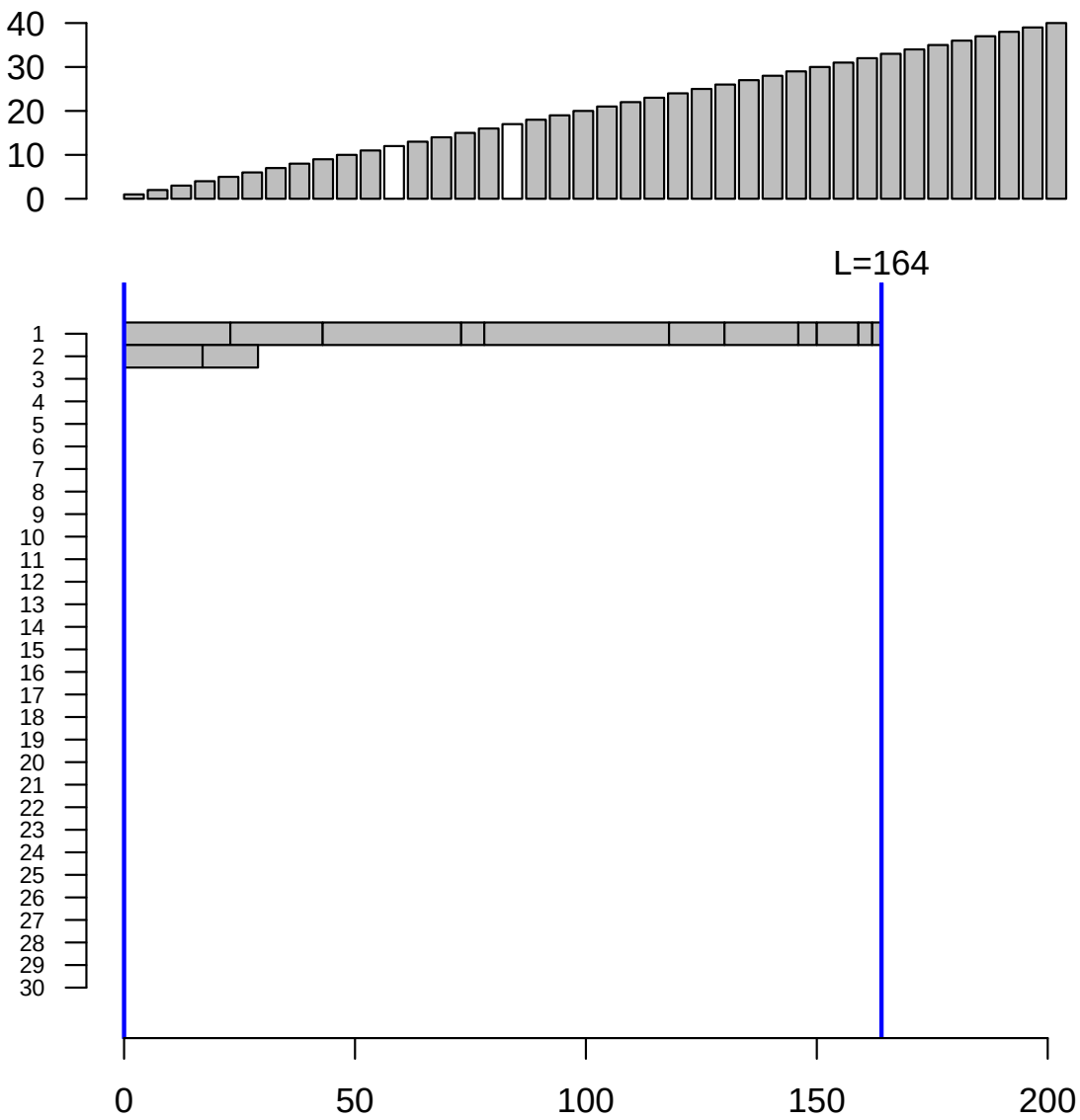
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



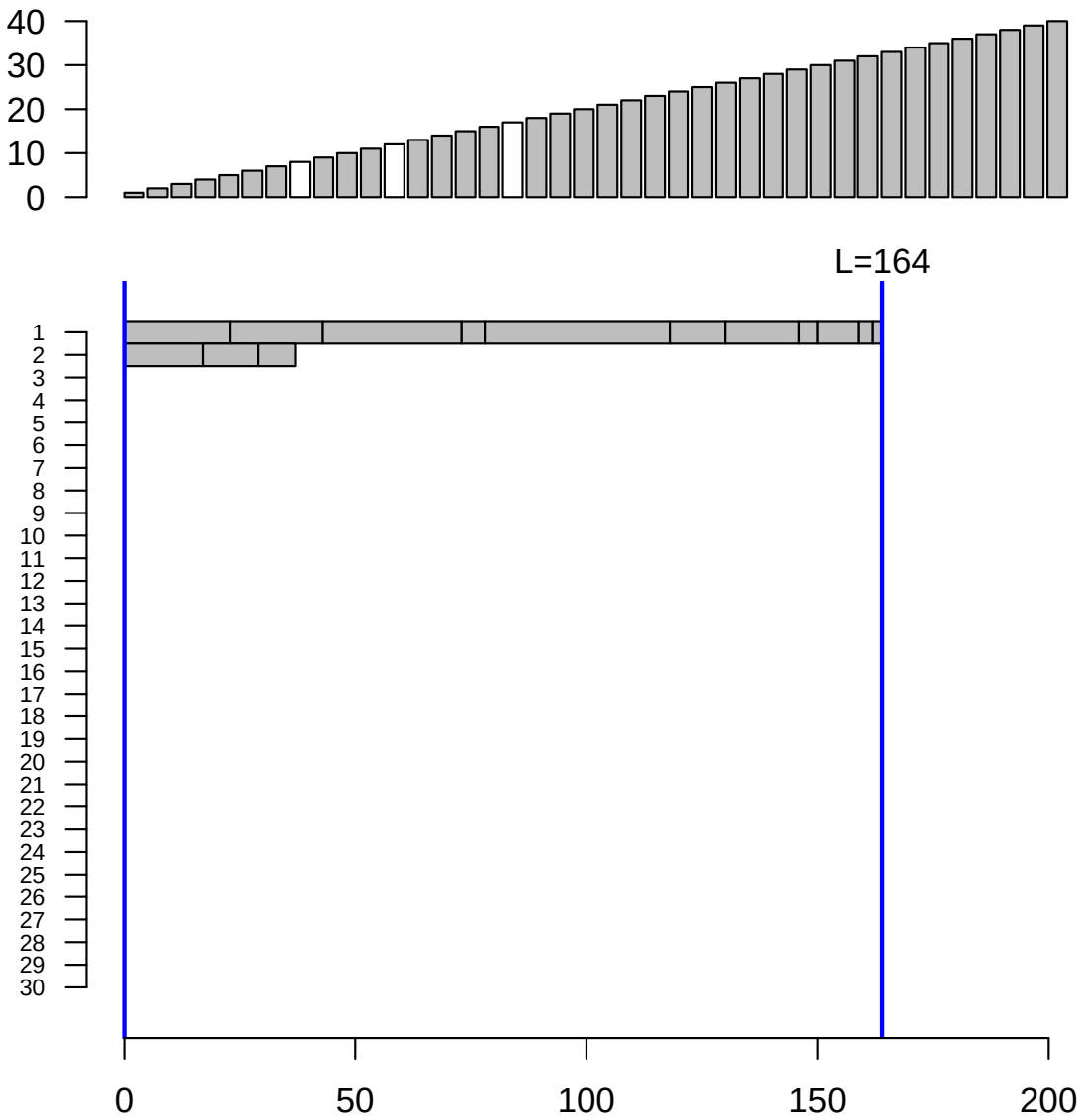
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



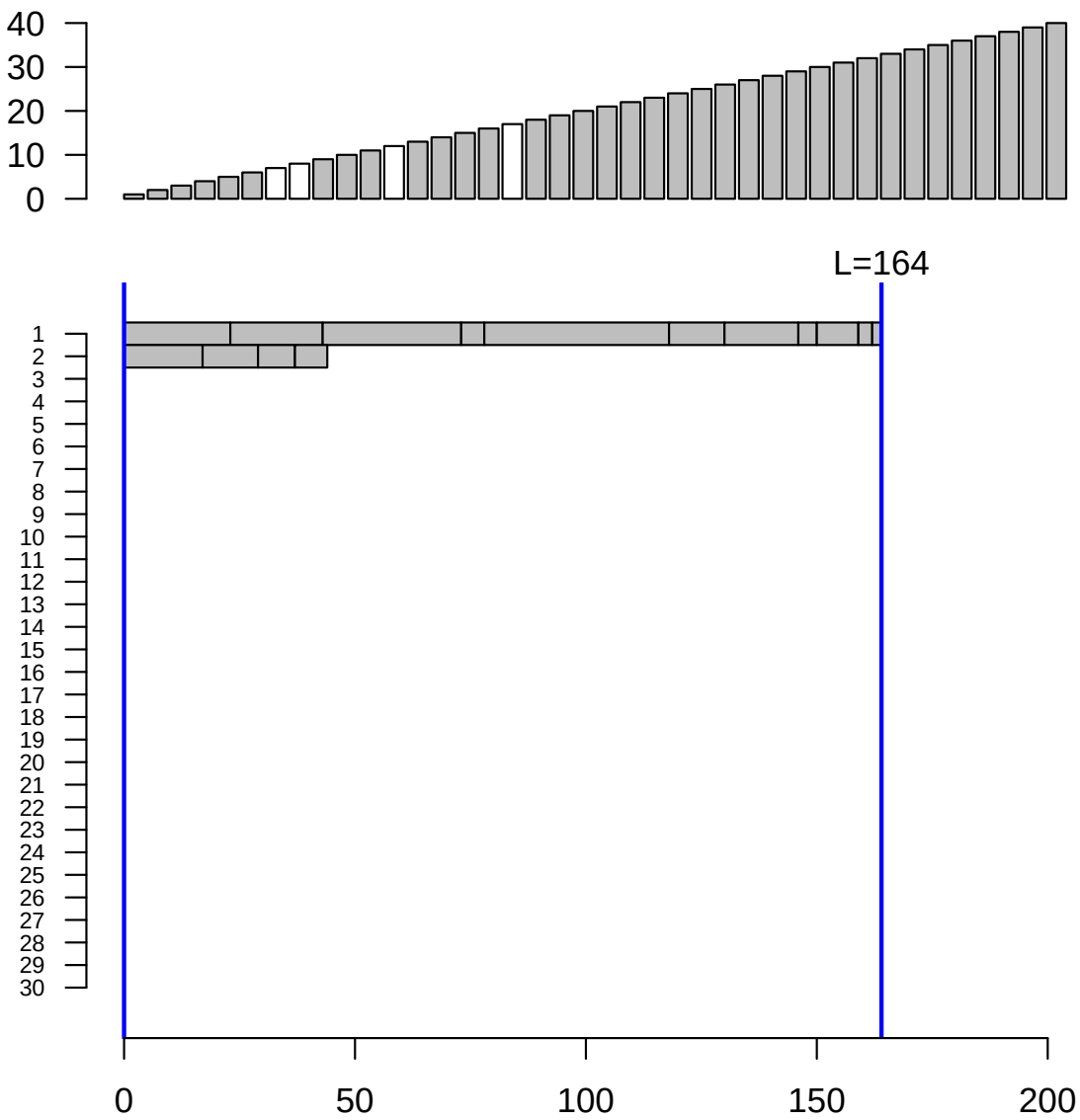
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



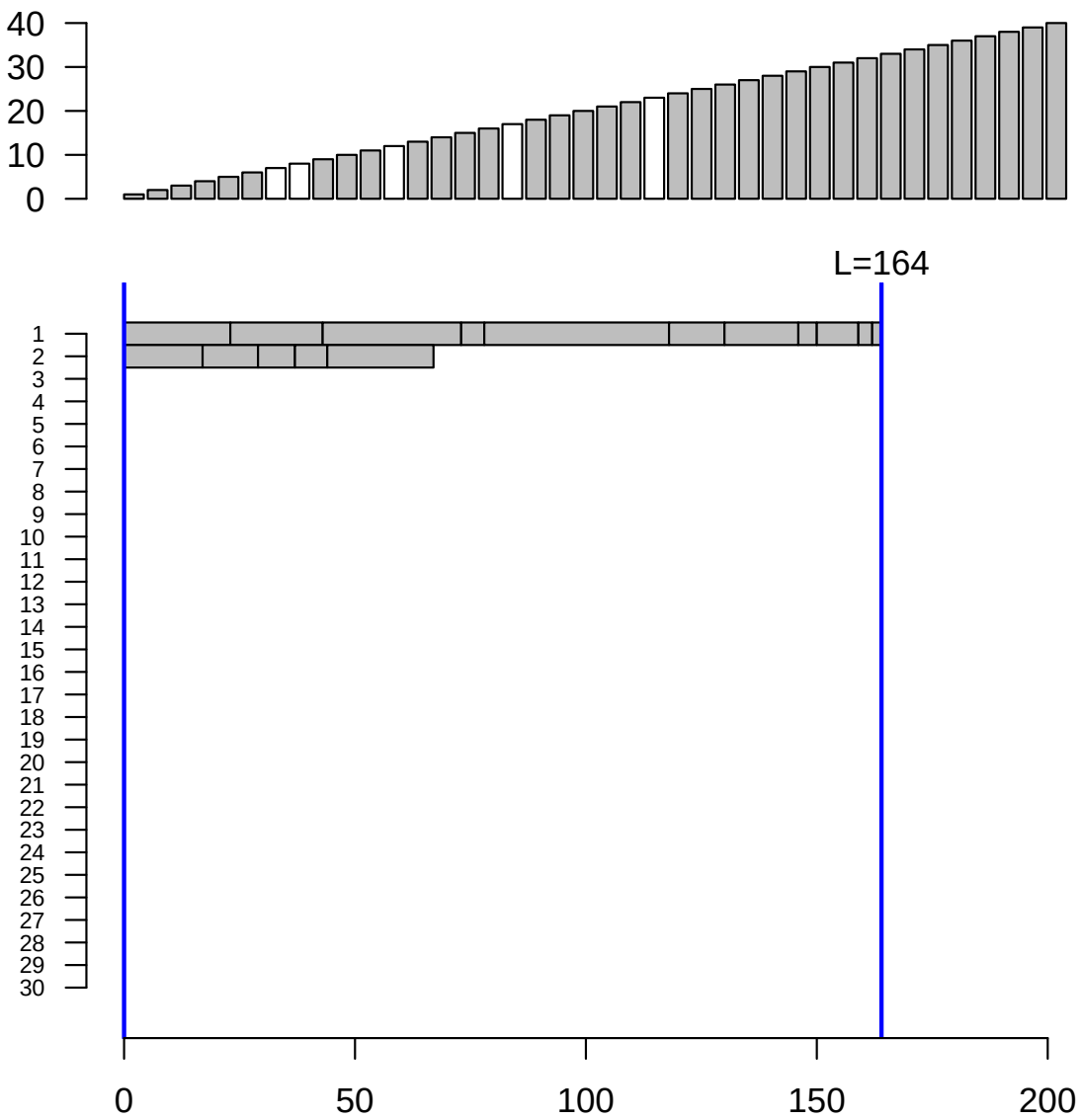
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



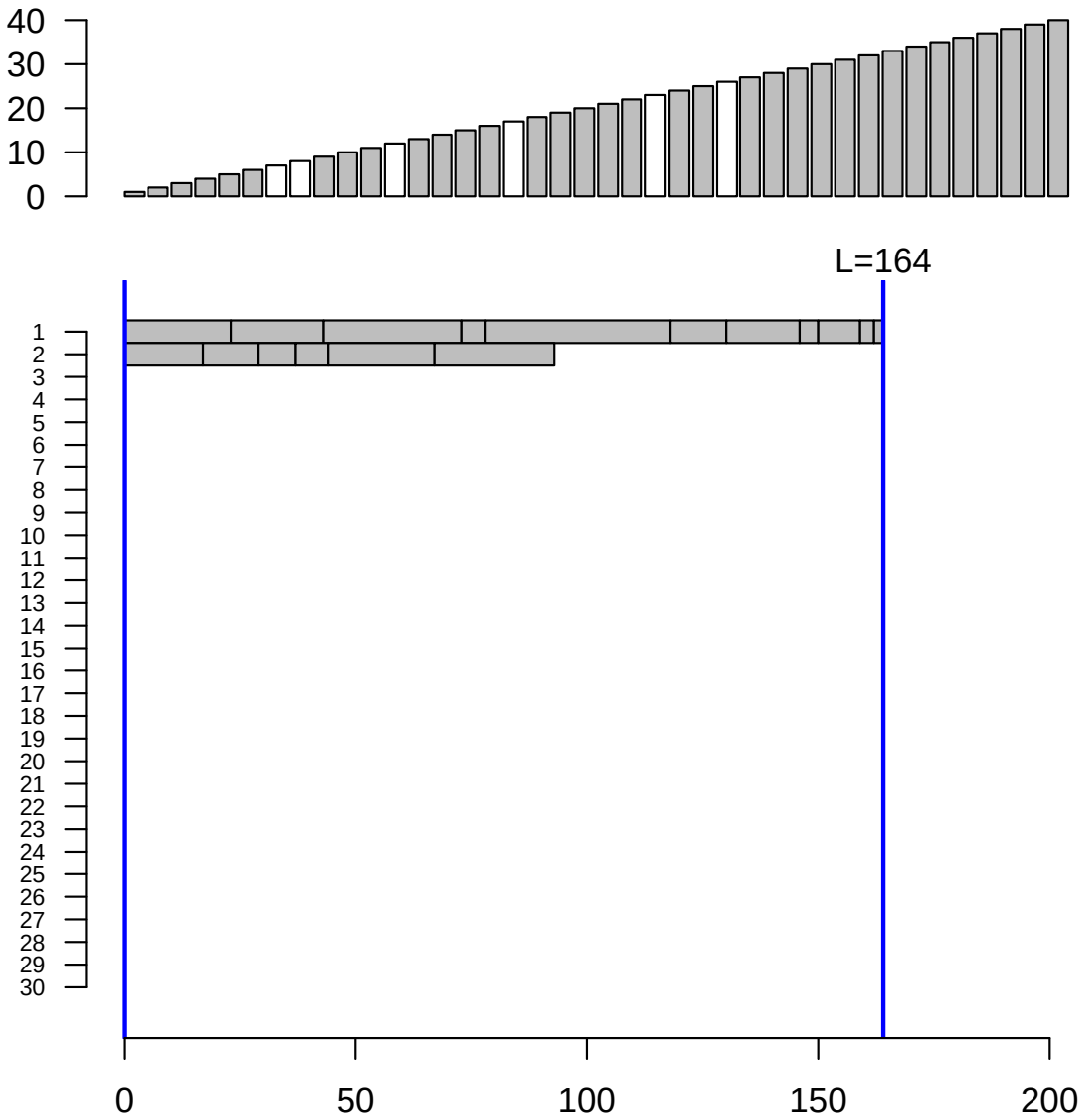
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



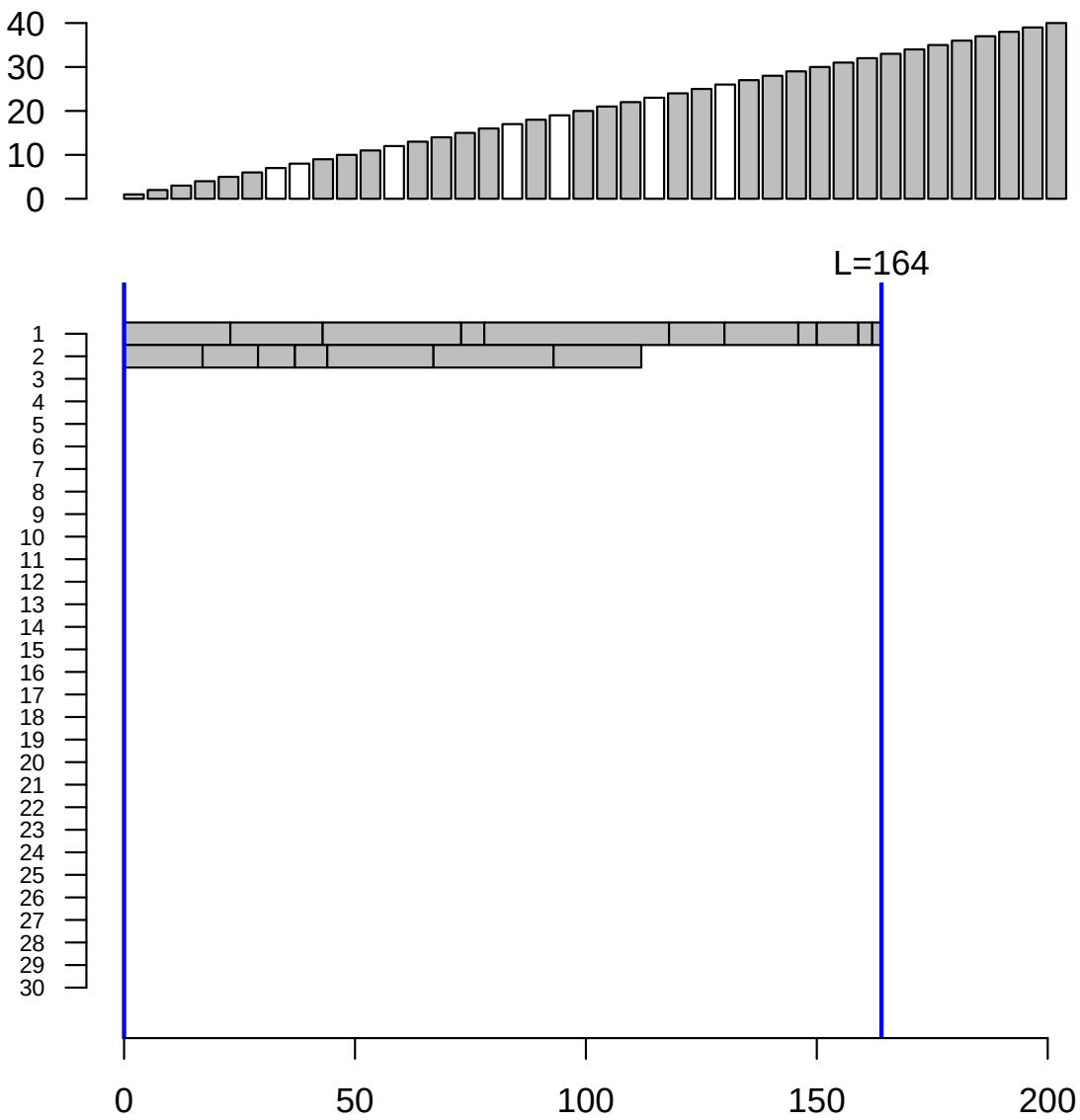
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



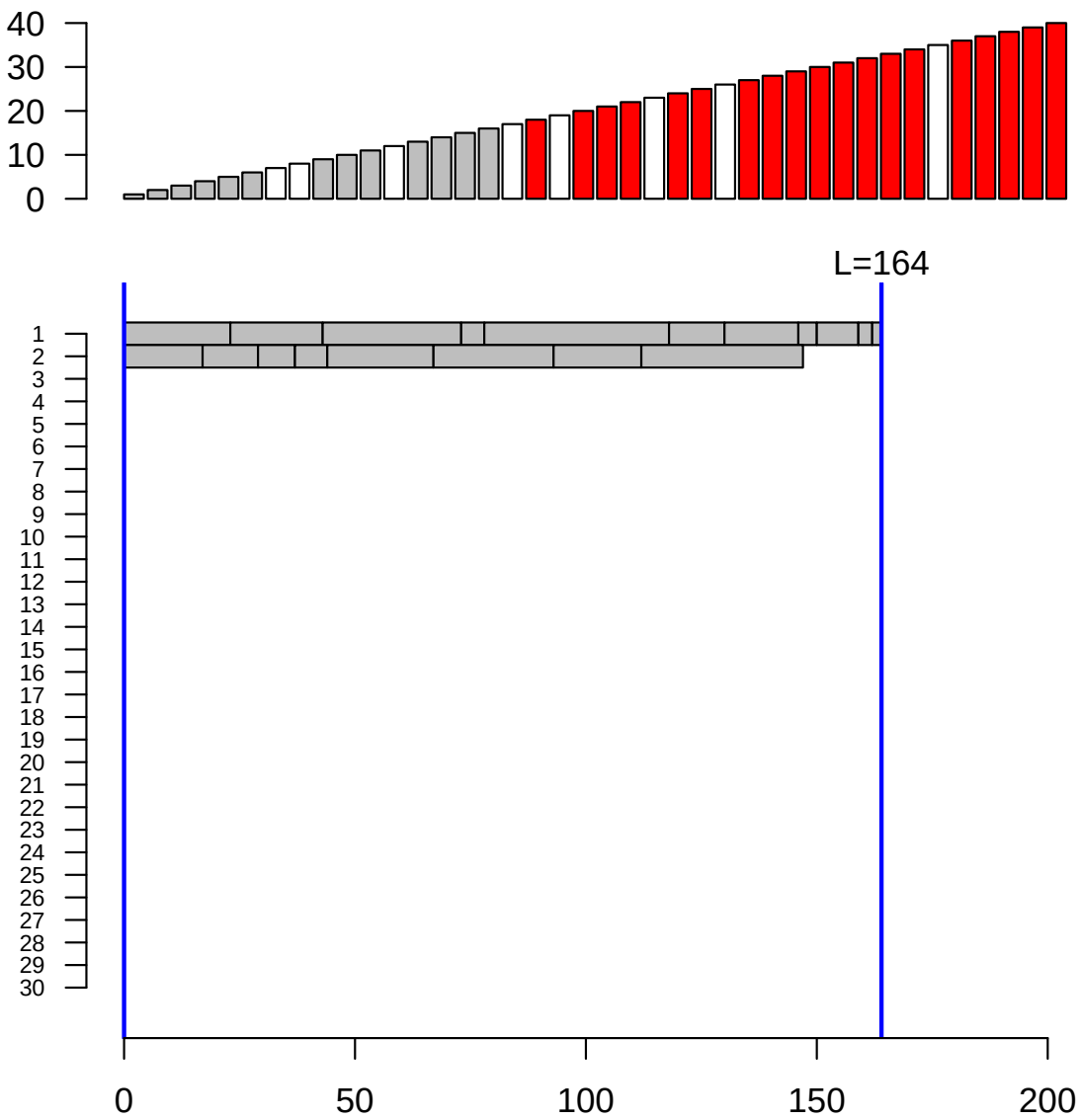
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



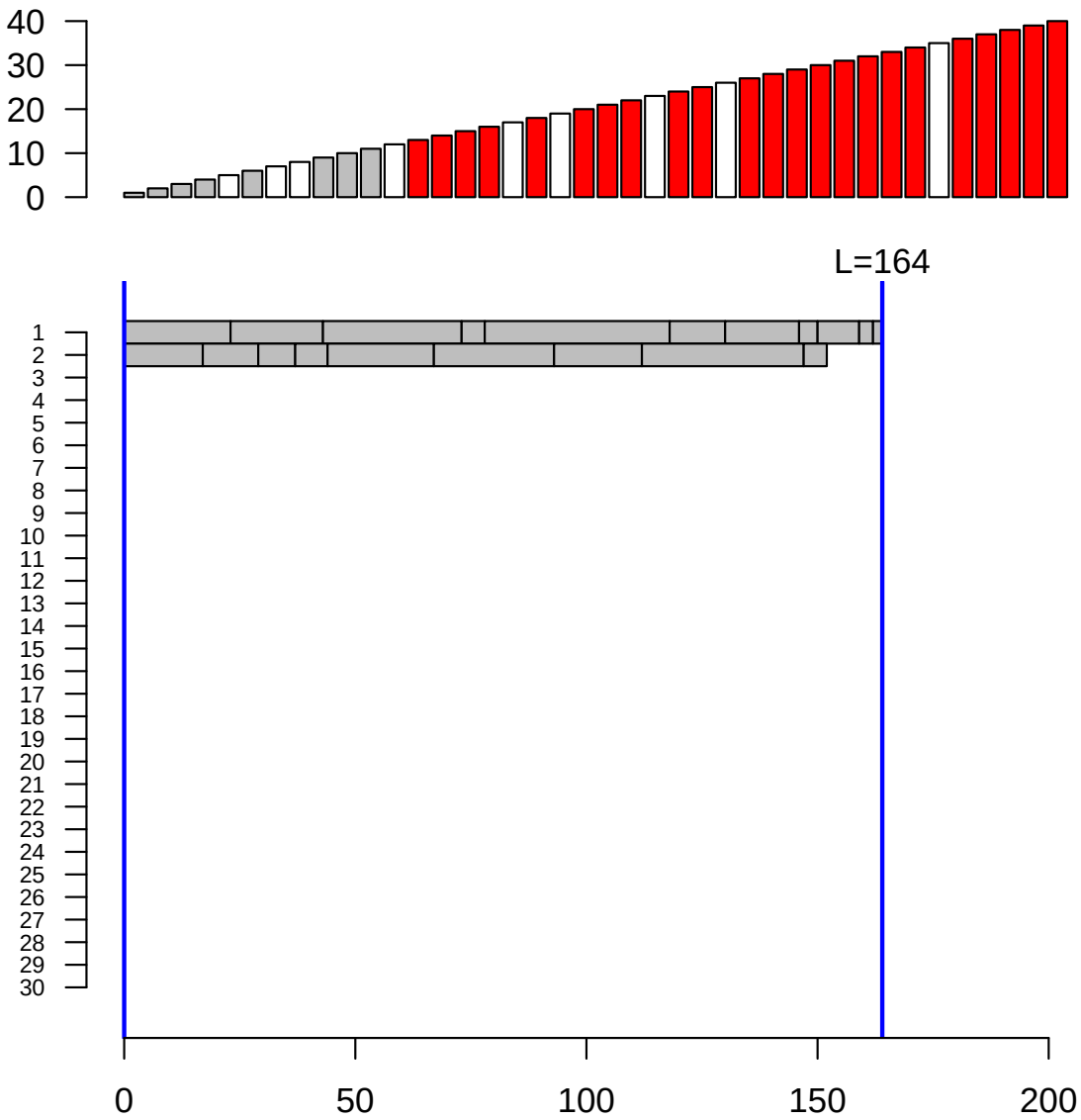
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



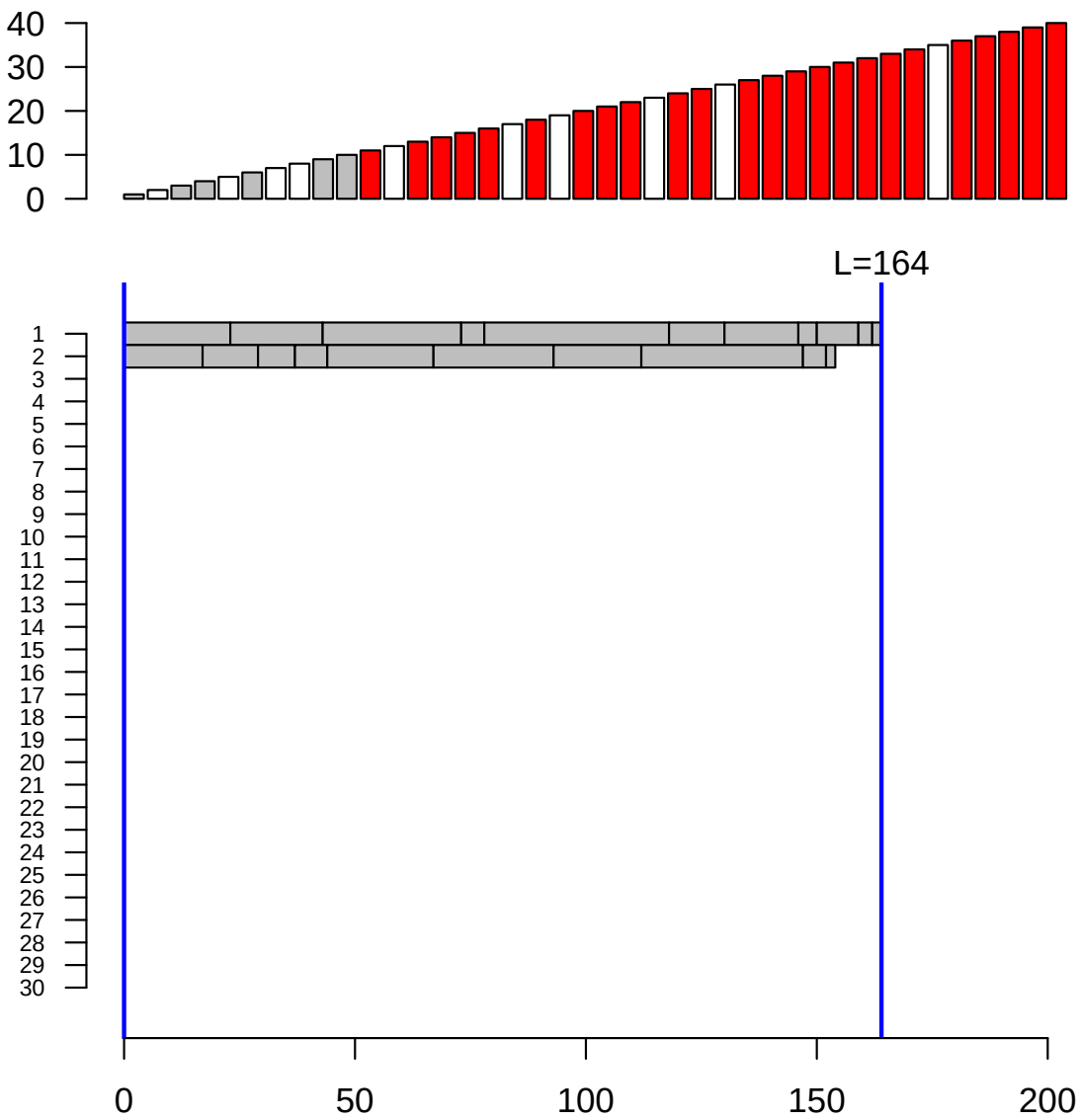
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



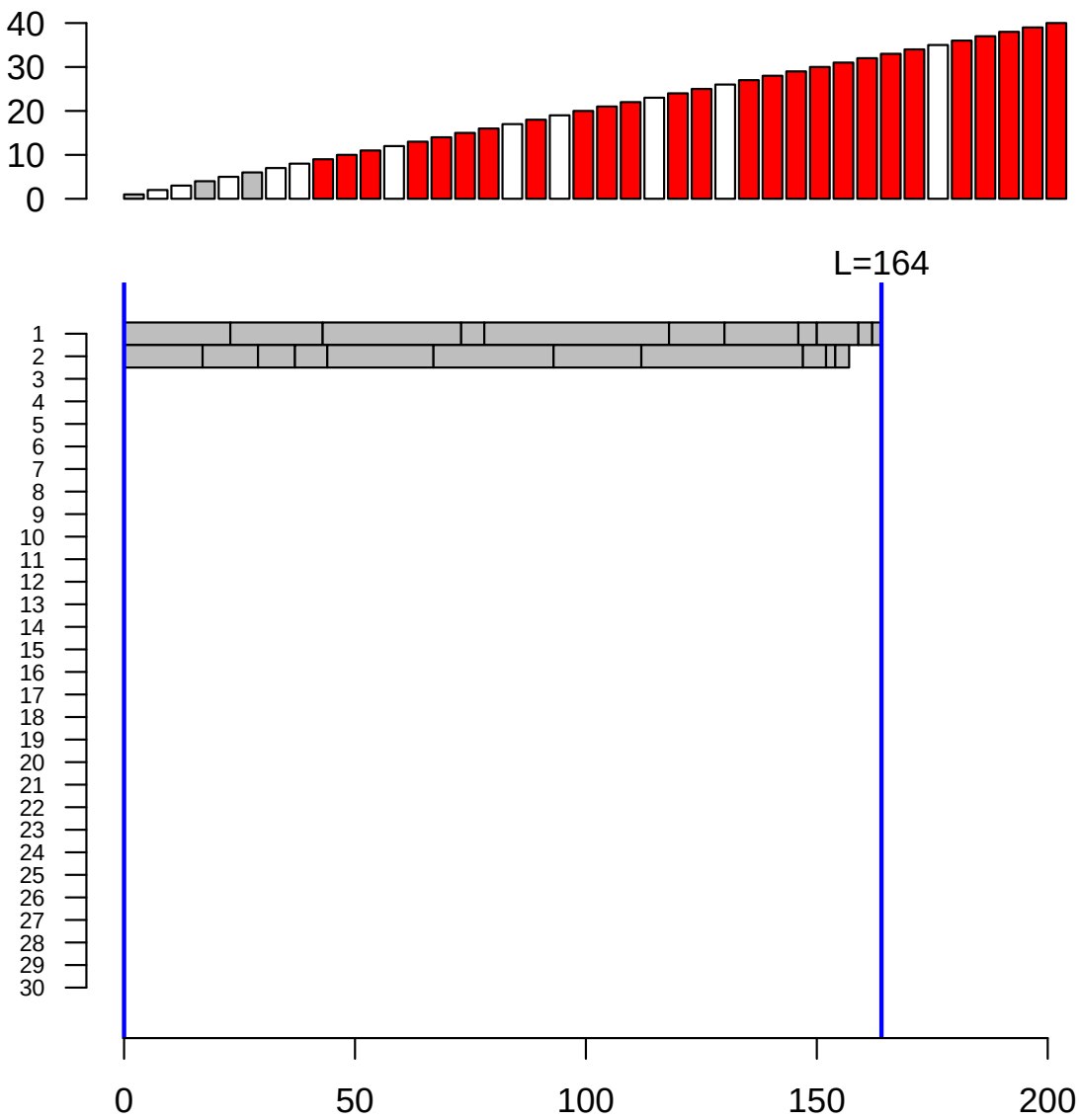
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



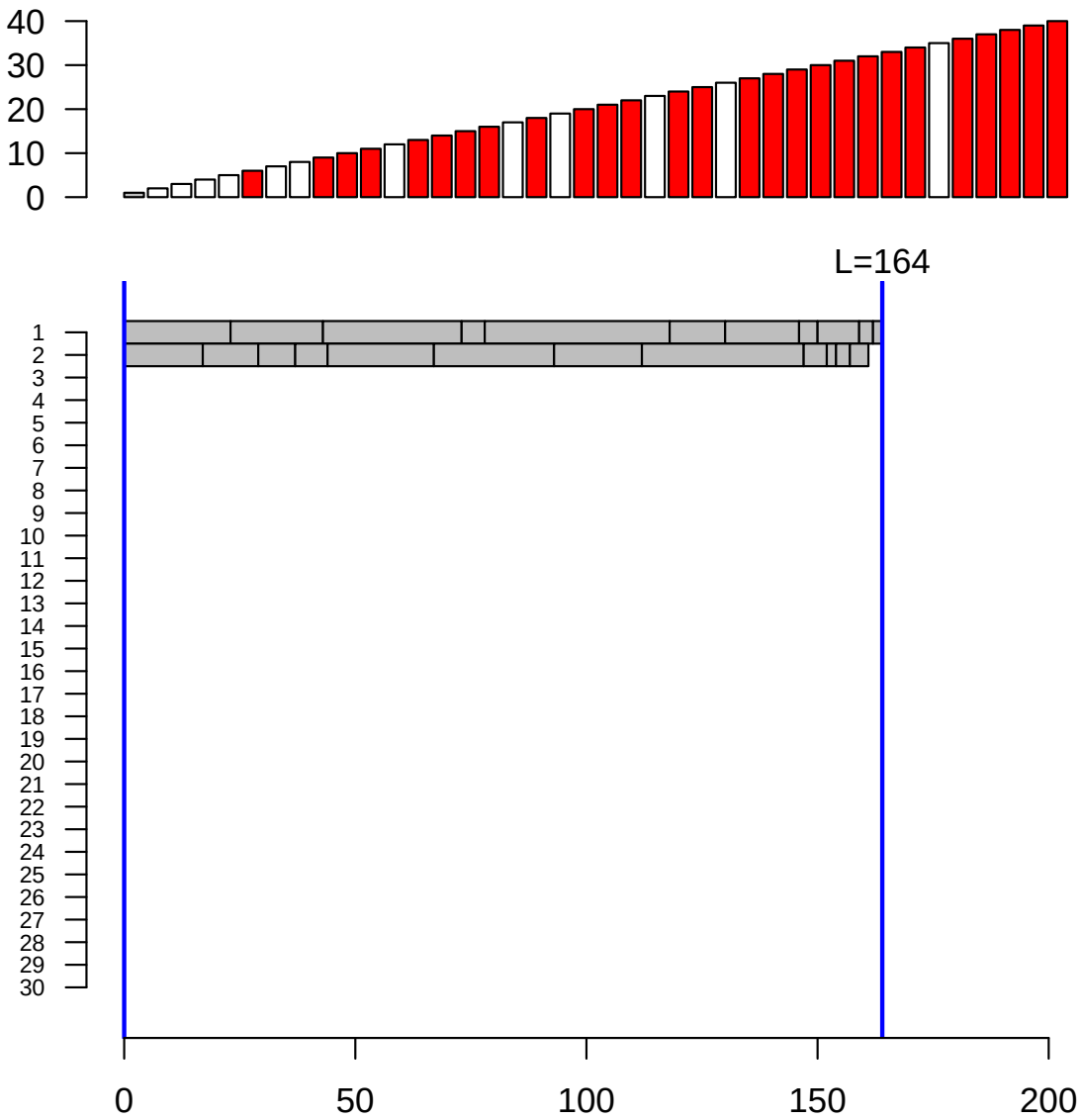
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



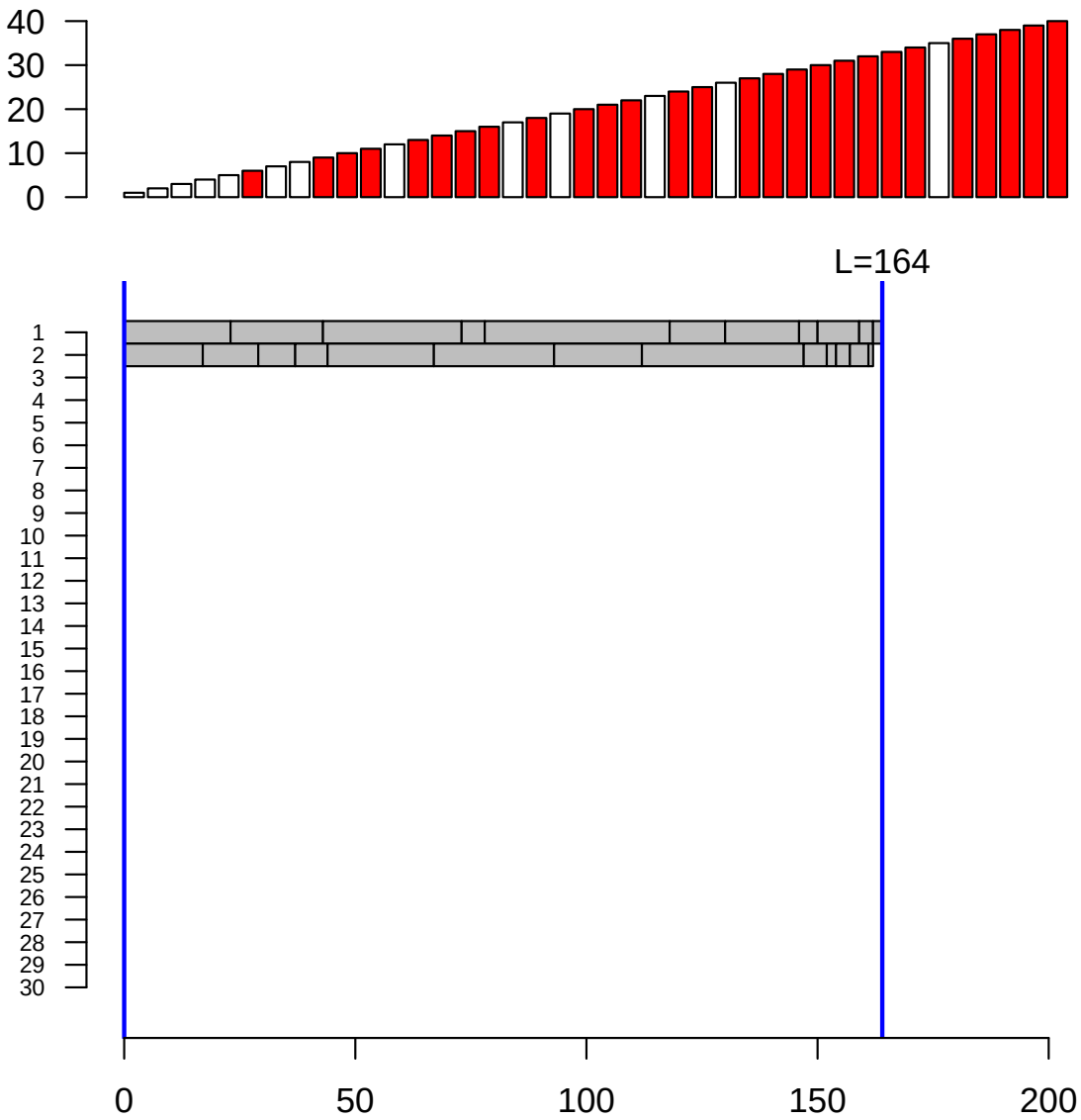
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



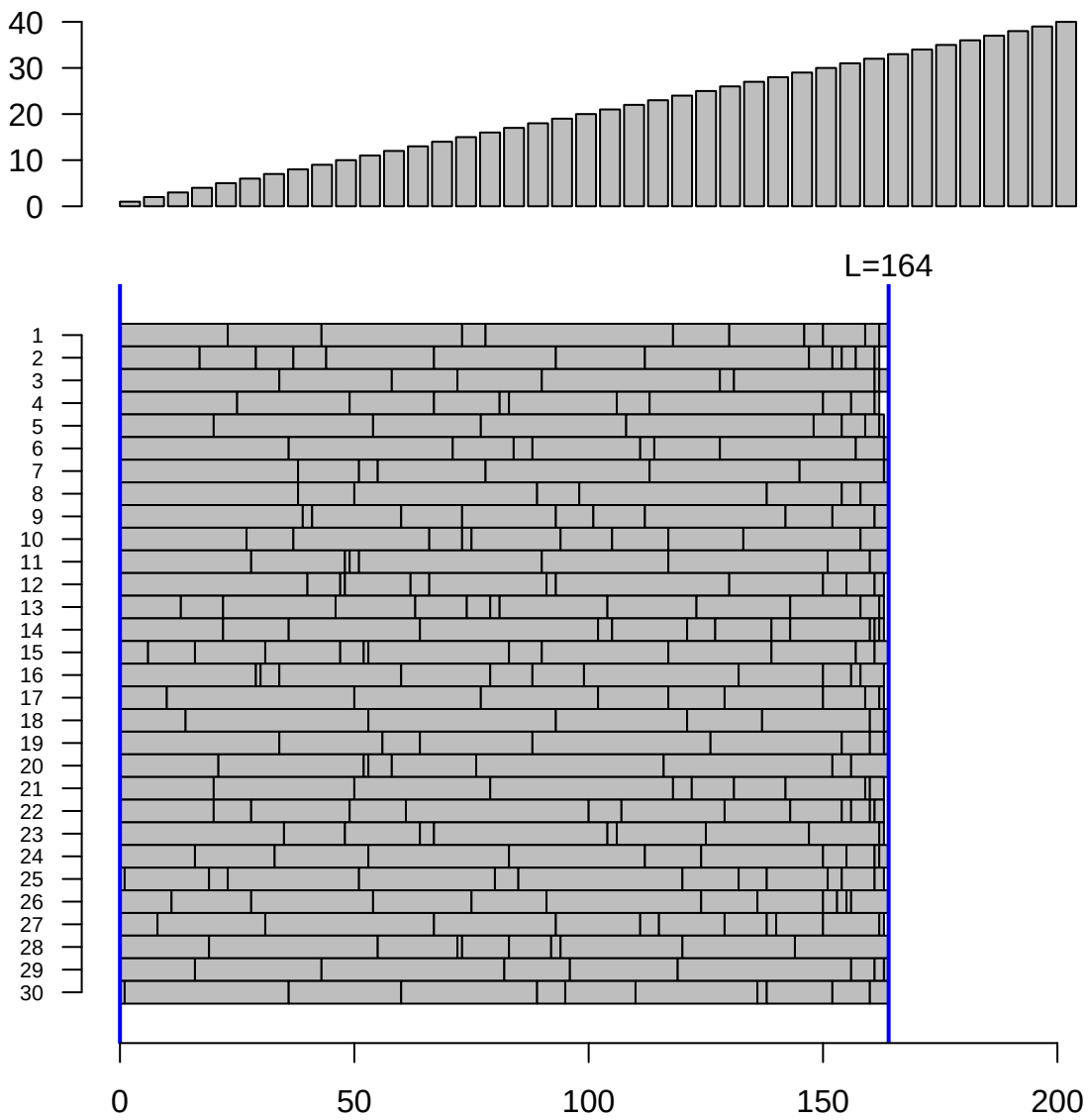
Example 7

How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



Example 7

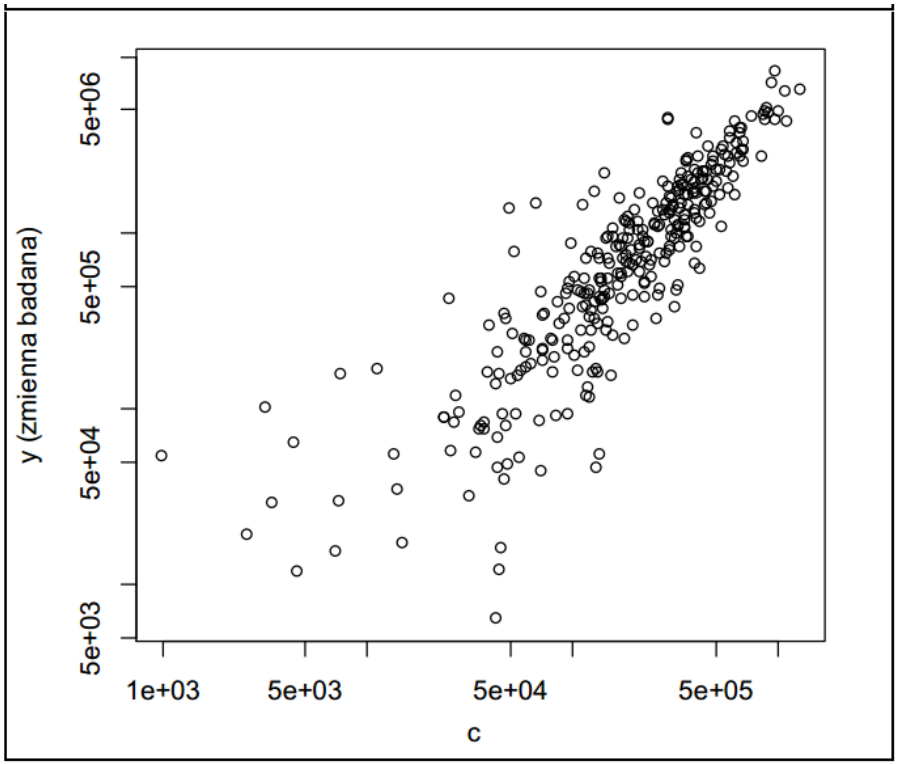
How the budget is spent ? (Greedy, $\mathbf{c} = [1, ..., 40]'$, $L = 164$)



Back to example 5

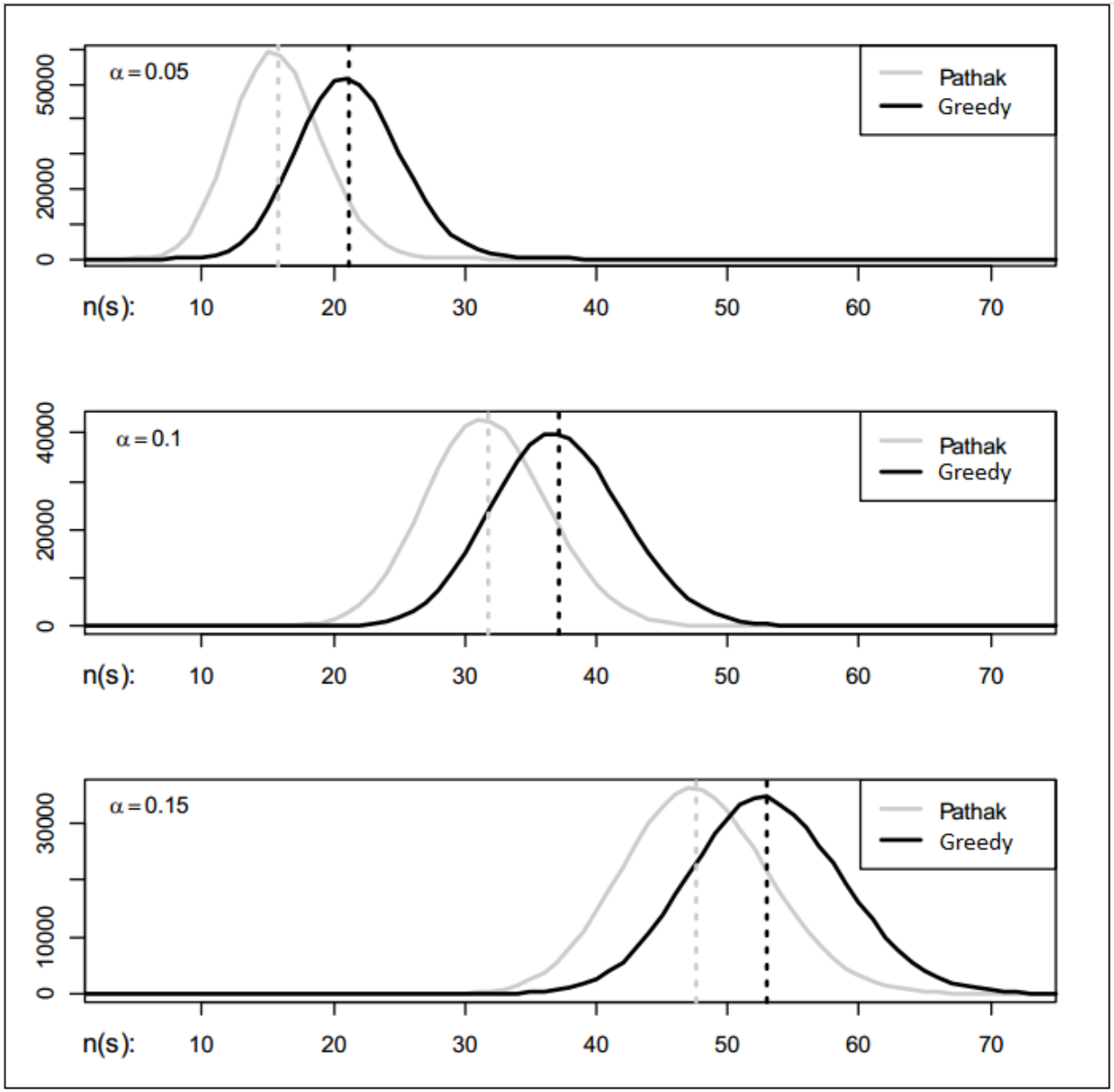
How are sample sizes distributed in this setup ?

- A population of $N = 319$ KZK-GOP communication lines
- Study variable: yearly number of passengers
- Sampling cost proportional to yearly vehicle miles driven
- Joint distribution of cost vs study variable



Back to example 5

Sample size distributions



Inclusion probabilities

Property 1

Individual first-order inclusion probabilities under greedy scheme are not lower than those under Pathak scheme

Property 2

Expected sample size under greedy scheme is not lower than under Pathak scheme

However:

- For both schemes probabilities are hard to calculate
- For both schemes cheaper units are chosen more often
- For the greedy scheme differences are more pronounced

How to estimate the mean ?

- For known $\pi_1, ..., \pi_N$, an unbiased estimator would be:

$$\overline{y}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i} \tag{12}$$

with the variance:

$$V(\overline{y}_{HT}) = \frac{1}{N^2} \sum_{i,j \in U} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \tag{13}$$

and its estimator:

$$\hat{V}(\overline{y}_{HT}) = \frac{1}{N^2} \sum_{i,j \in s} \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \tag{14}$$

- However $\pi_1, ..., \pi_N$ are not known

Fattorini's (2006,2009) solution
simulate them

- As assumed, individual costs are known
- Draw R independent replications of s (same scheme)
- Calculate counts of occurences $k_1, ..., k_N$
- Compute *empirical inclusion probabilities* $\hat{\pi}_1, ..., \hat{\pi}_N$
- Plug them into estimator:

$$\overline{y}_{EHT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\hat{\pi}_i} \tag{15}$$

How to estimate inclusion probabilities ?
Some challenges ...

- Estimates must be positive for $\overline{y}_{EHT}$'s moments to be finite
- Stopping rule providing desired accuracy is needed
- Inclusion of units is not independent
- True probabilities often very small

How to estimate inclusion probabilities ?

Some ideas ...

Thompson & Wu (2008)

$$\hat{\pi}_i = \frac{k_i}{R} \tag{16}$$

Fattorini (2006,2009):

$$\hat{\pi}_{iF} = \frac{k_i + 1}{R + 1} \tag{17}$$

Gamrot (2012): use isotonic regression

Gamrot (2013): use RML

Gamrot (2013): use kernel smoothing

Gamrot (2014): use sampling weight posterior mode/mean

Assessing properties of mean estimators ?
(design-and-randomization-based properties)

Fattorini's results for $\hat{\pi}_{iF}$:

$$ARB(\overline{y}_{EHT}) = \frac{|E(\overline{y}_{EHT}) - \overline{y}|}{\overline{y}} \leq \frac{1}{(R+2)\pi_-}$$

$$\frac{|MSE(\overline{y}_{EHT}) - V(\overline{y}_{HT})|}{V(\overline{y}_{HT})} \leq \frac{9}{(R+2)\pi_-}\left(1 + \frac{\overline{y}^2}{V(\overline{y}_{HT})}\right)$$

But what is exact (non-relative) bias ?

And what about other estimators ?

Perhaps a simulation ?

A naive approach:

- Find some complete data set with known y-values
- Repeat independently $H >> 1$ times following steps:
  1. Draw s from U
  2. 'Observe' study variable in s
  3. Generate $R >> 1$ replications $s'_1, ..., s'_R$ of s
  4. Count occurrences of sampled units
  5. Calculate estimates of inclusion probabilities
  6. Plug everything into $\overline{y}_{EHT}$
- Examine the empirical distribution of $\overline{y}_{EHT}$

Perhaps a simulation ?

The naive approach:

- simulates repeated sampling
- is easy to justify intuitively
- is very simple

However:

- the number of sample draws is $H \cdot (R + 1)$
- when $H, R >> 1$, it may be infeasible
- simulating for $R = R_1, ..., R_J$, it is even harder

A better alternative

- Consider a random vector $\mathbf{Z} = [z_1, ..., z_N]'$ where

$$z_i = \begin{cases} 0 & for \quad i \in s \\ 1 & for \quad i \notin s \end{cases}$$

- Consider a random vector $\mathbf{W} = [w_1, ..., w_N]'$ where:

$$w_i = \frac{y_i}{N\hat{\pi}_i}$$

- **W** and **Z** are independent
- Components within **W** are not
- Components within **Z** are not
- The H-T estimator is expressed as:

$$\overline{y}_{EHT} = \mathbf{W}'\mathbf{Z}$$

A better alternative
for assessing the bias

- Consequently

$$E(\overline{y}_{EHT}) = E(\mathbf{W}')E(\mathbf{Z})$$

- Simulate independently:
  - m realizations $\mathbf{W}_1, ..., \mathbf{W}_m$ of $\mathbf{W}$
  - n realizations $\mathbf{Z}_1, ..., \mathbf{Z}_n$ of $\mathbf{Z}$

- This requires $m \cdot R + n$ sample draws
- Denote $\mathbf{W}_i = [w_{i1}, ..., w_{iN}]', \quad \mathbf{Z}_i = [z_{i1}, ..., z_{iN}]'$

A better alternative
for assessing the bias

- Calculate means

$$\overline{\mathbf{W}}_m = [\overline{w_1}, ..., \overline{w_N}] = \frac{1}{m} \sum_{i=1}^{m} \mathbf{W}_i$$

$$\overline{\mathbf{Z}}_n = [\overline{z_1}, ..., \overline{z_N}] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i$$

- They are respectively unbiased for $E(\mathbf{W})$ and $E(\mathbf{Z})$
- Hence the estimator:

$$\hat{E}(\overline{y}_{EHT}) = \overline{\mathbf{W}}'_m \overline{\mathbf{Z}}_n$$

- is unbiased for $E(\overline{y}_{EHT})$

A better alternative
for assessing the bias

- Let $\mathbf{C}_W$ and $\mathbf{C}_Z$ are covariance matrices of $\mathbf{W}$ and $\mathbf{Z}$
- The estimator variance is:

$$V(\hat{E}(\overline{y}_{EHT})) = \frac{1}{mn}G_1 + \frac{1}{n}G_2 + \frac{1}{m}G_3$$

where

$$G_1 = tr\ \mathbf{C}_W\mathbf{C}_Z, \quad G_2 = E(\mathbf{W}')\mathbf{C}_Z E(\mathbf{W}) \quad G_3 = E(\mathbf{Z}')\mathbf{C}_W E(\mathbf{Z})$$

- How to set m, n minimizing the variance ?

A better alternative
for assessing the bias

Covariance matrices are unbiasedly estimated by:

$$\hat{\mathbf{C}}_W = \begin{bmatrix} \hat{c}_{w_1,w_1} & \cdots & \hat{c}_{w_1,w_N} \\ \vdots & \ddots & \vdots \\ \hat{c}_{w_N,w_1} & \cdots & \hat{c}_{w_N,w_N} \end{bmatrix} \qquad \hat{\mathbf{C}}_Z = \begin{bmatrix} \hat{c}_{z_1,z_1} & \cdots & \hat{c}_{z_1,z_N} \\ \vdots & \ddots & \vdots \\ \hat{c}_{z_N,z_1} & \cdots & \hat{c}_{z_N,z_N} \end{bmatrix}$$

where

$$\hat{c}_{w_g,w_h} = \frac{1}{m-1} \sum_{i=1}^{m} (w_{gi} - \overline{w}_g)(w_{hi} - \overline{w}_h)$$

$$\hat{c}_{z_g,z_h} = \frac{1}{n-1} \sum_{i=1}^{n} (z_{gi} - \overline{z}_g)(z_{hi} - \overline{z}_h)$$

for $g, h \in U$

A better alternative
for assessing the bias

Constants G_1, G_2, G_3 are unbiasedly estimated by:

$$\hat{G}_1 = tr\ \hat{\mathbf{C}}_W \hat{\mathbf{C}}_Z$$

$$\hat{G}_2 = \overline{\mathbf{W}}'_m \hat{\mathbf{C}}_Z \overline{\mathbf{W}}_m - \frac{1}{m} tr\ \hat{\mathbf{C}}_Z \hat{\mathbf{C}}_W$$

$$\hat{G}_3 = \overline{\mathbf{Z}}'_n \hat{\mathbf{C}}_W \overline{\mathbf{Z}}_n - \frac{1}{n} tr\ \hat{\mathbf{C}}_W \hat{\mathbf{C}}_Z$$

The variance of $\hat{E}(\overline{y}_{EHT})$ is estimated without bias by:

$$\hat{V}(\hat{E}(\overline{y}_{EHT})) = \nu(m,n) = \frac{1}{mn}\hat{G}_1 + \frac{1}{n}\hat{G}_2 + \frac{1}{m}\hat{G}_3$$

An interesting special case

For $m = n = 1$ we have:

$$\hat{E}(\overline{y}_{EHT}) = \overline{y}_{EHT}$$

And consequently:

$$V(\overline{y}_{EHT}) = G_1 + G_2 + G_3$$

estimated without bias by

$$\hat{V}(\overline{y}_{EHT}) = \hat{G}_1 + \hat{G}_2 + \hat{G}_3$$

Back to setting simulation parameters

- From symmetry an non-negative defininiteness of covariance matrices we have $G_1, G_2, G_3 \geq 0$
- So we should also get $\hat{G}_1, \hat{G}_2, \hat{G}_3 \geq 0$
- Hence

$$\nu(m,n) = \frac{1}{mn}\hat{G}_1 + \frac{1}{n}\hat{G}_2 + \frac{1}{m}\hat{G}_3$$

is a convex function of m, n

A simulation procedure

- Let the simulation time be:

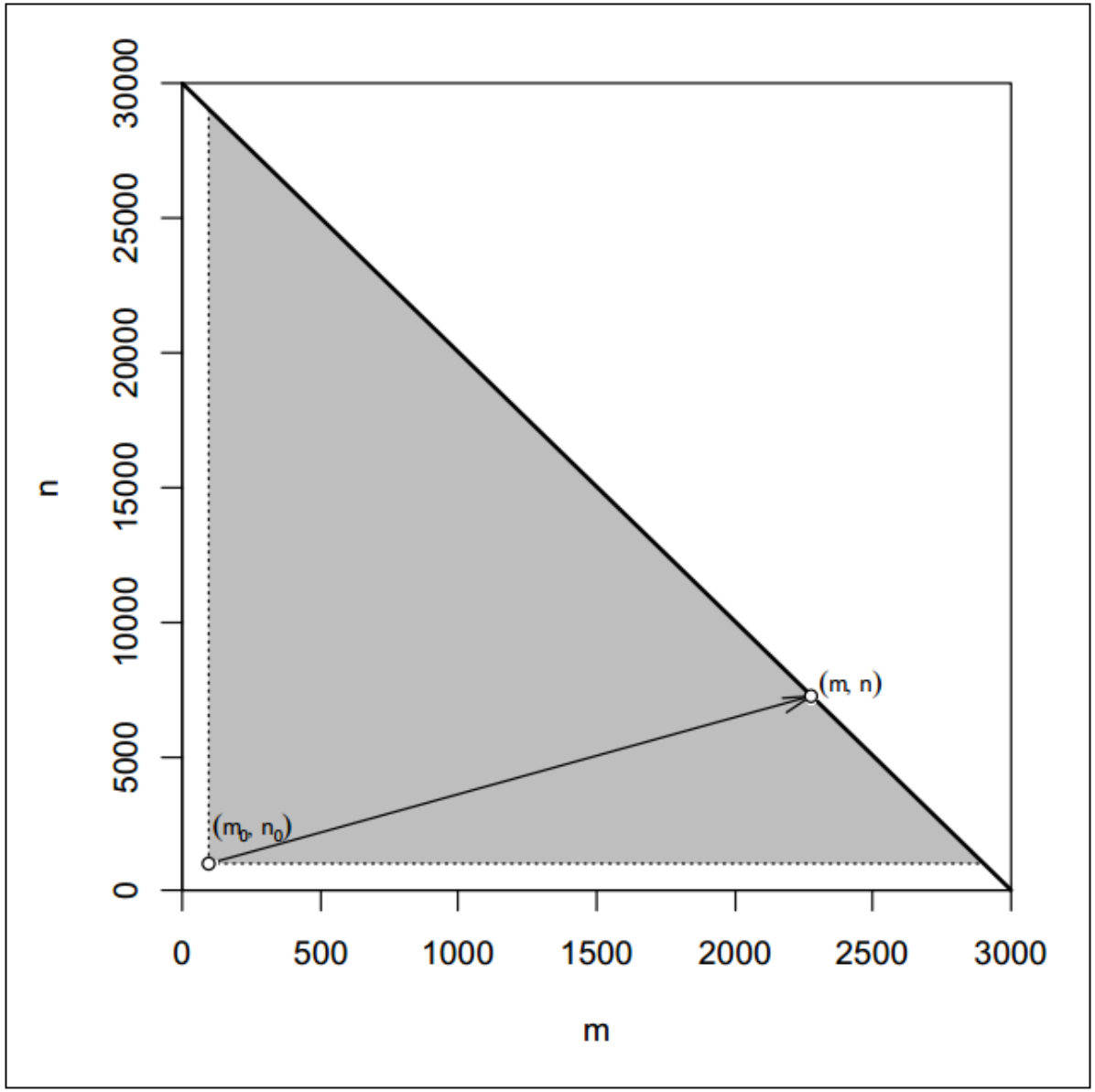
$$T(m,n) = mt_W + nt_Z$$

- Generate m_0 realizations of **W** and n_0 realizations of **Z**
- Estimate G_1, G_2, G_3, t_W, t_Z
- For a limited simulation time T_0 solve:

$$\begin{cases} \nu(m,n) \to min \\ T(m,n) \leq T_0 \\ m \geq m_0 \\ n \geq n_0 \end{cases}$$

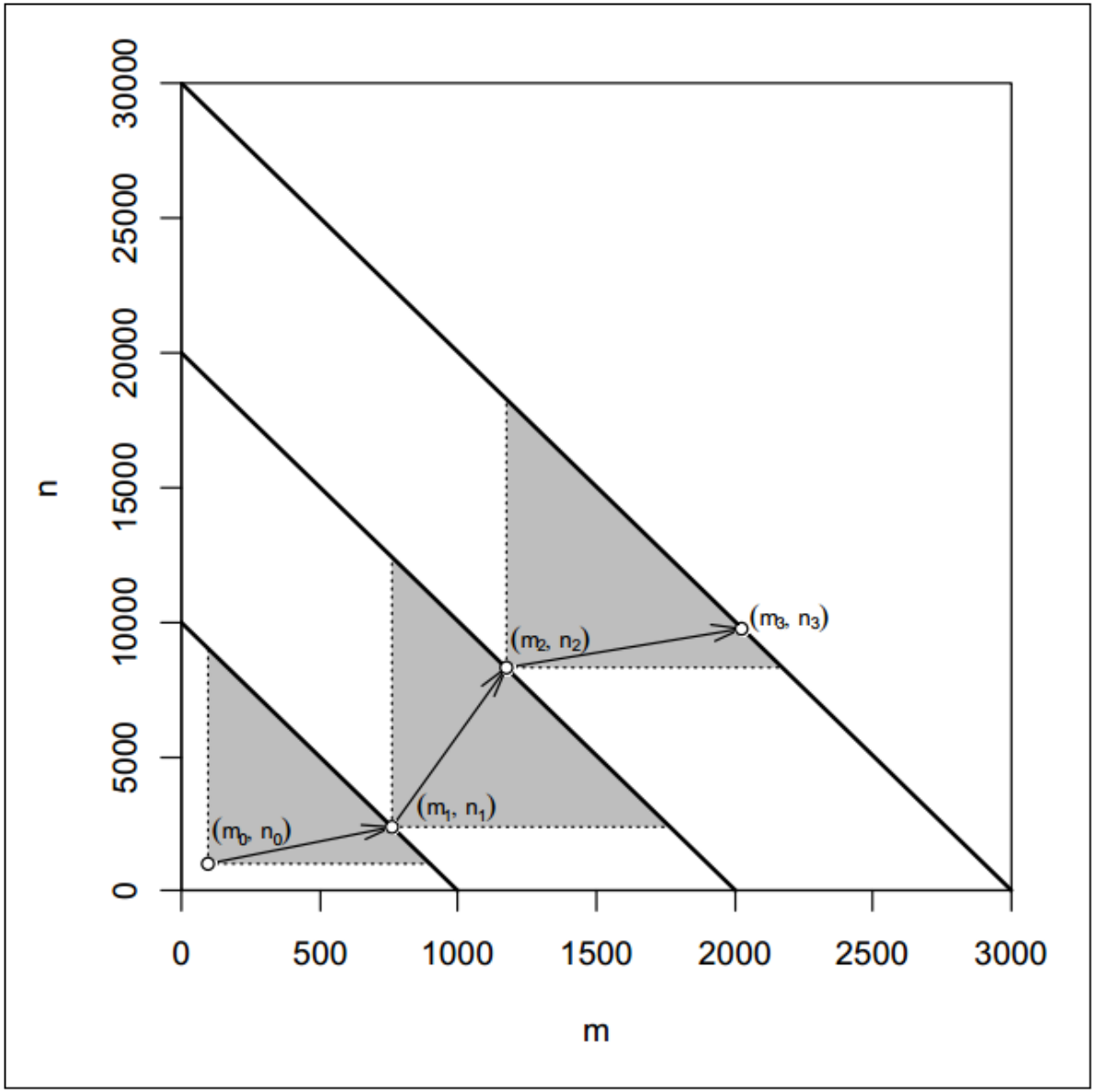
- Generate remaining $m - m_0$, $n - n_0$ realizations of **W**, **Z**
- Estimate the expectation of $\overline{y}_{EHT}$

A simulation procedure



A perhaps better simulation procedure

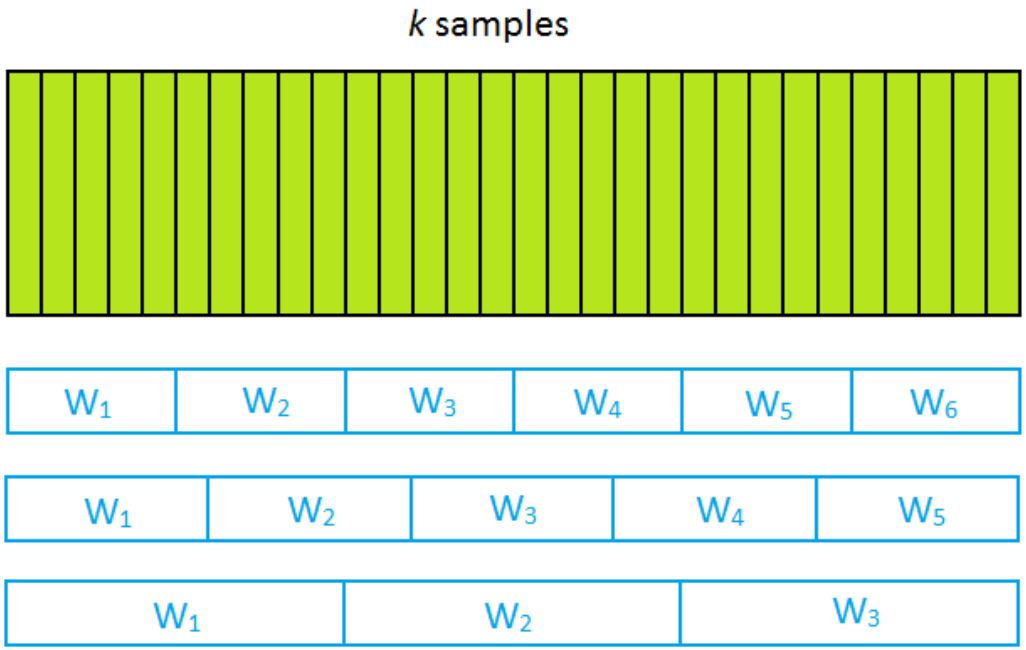
Sequence of time limits and optimization problems



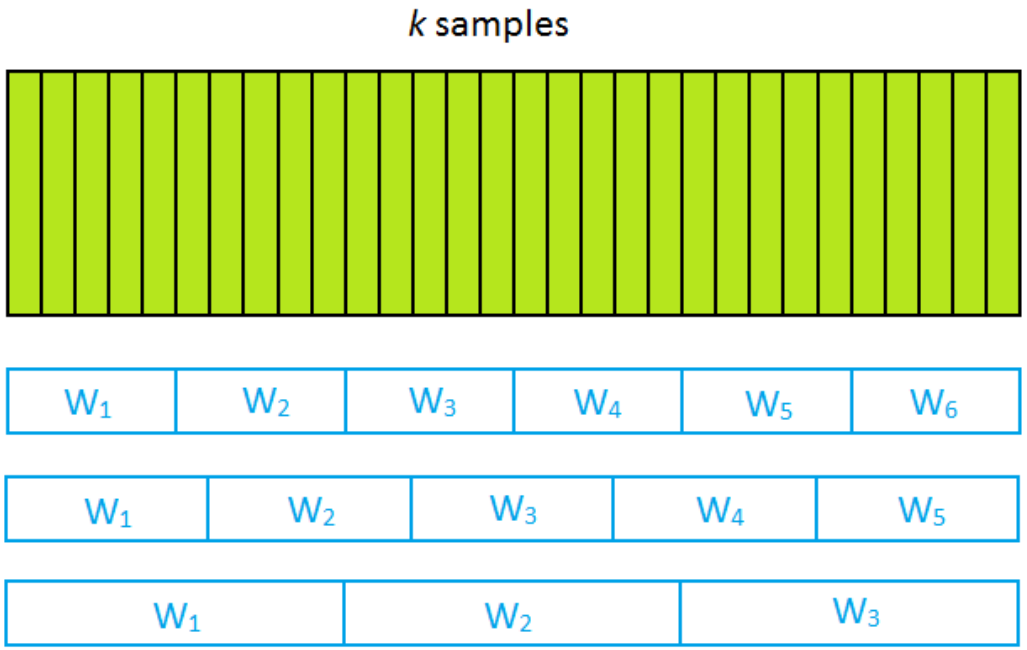
One more problem
worth solving

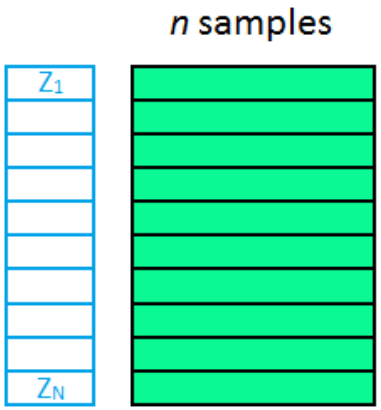
- How to assess estimator properties for $R_1, ..., R_J$?
- Do we need to repeat simulation J times ?
- Having $k = 10^7$ sample realizations one may compute:
  - 5000 independent **W**-values for $R = 2000$
  - 2500 independent **W**-values for $R = 4000$
  - 1666 independent **W**-values for $R = 6000$
  - 1250 independent **W**-values for $R = 8000$
  - 1000 independent **W**-values for $R = 10000$
- Generating of sample realizations for **W** takes time
- Recycling them would save time
- Recycling some n realizations of **Z** saves time as well

Simulating for multiple R's

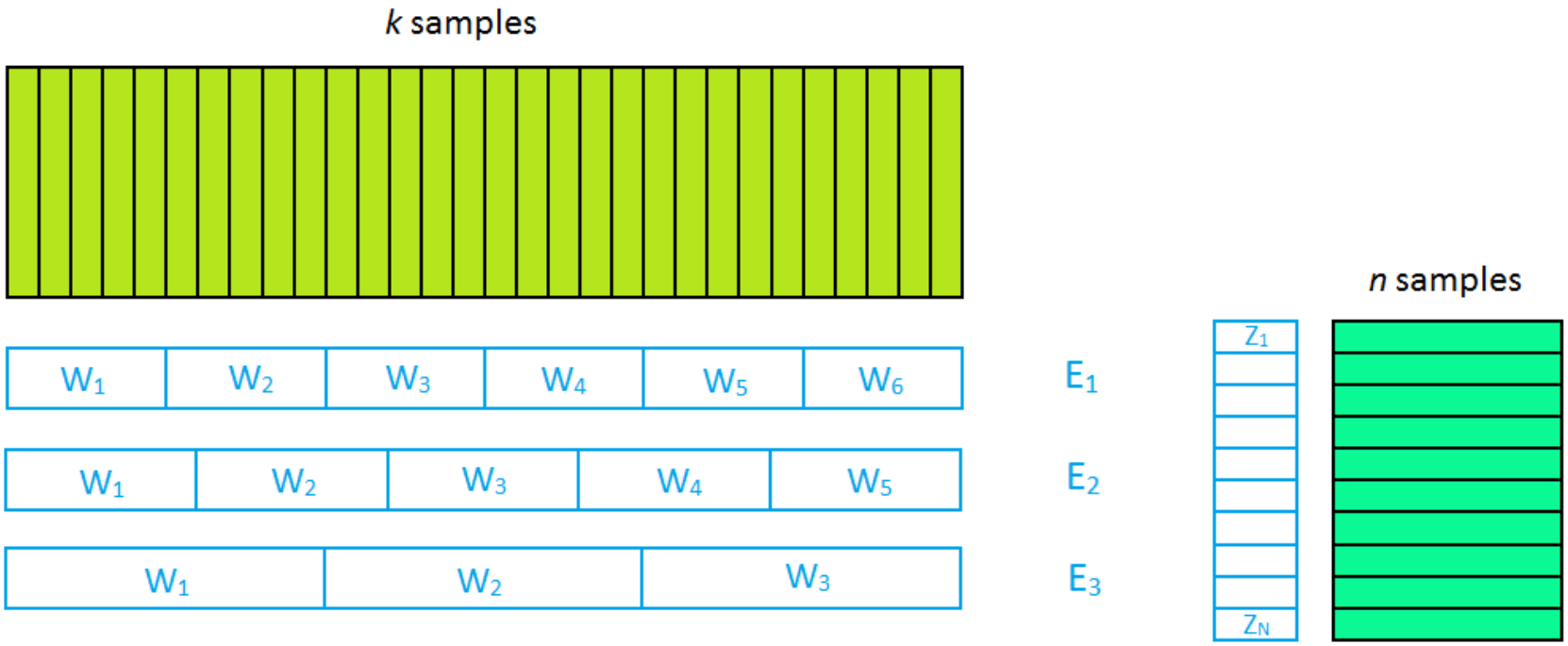


Simulating for multiple R's

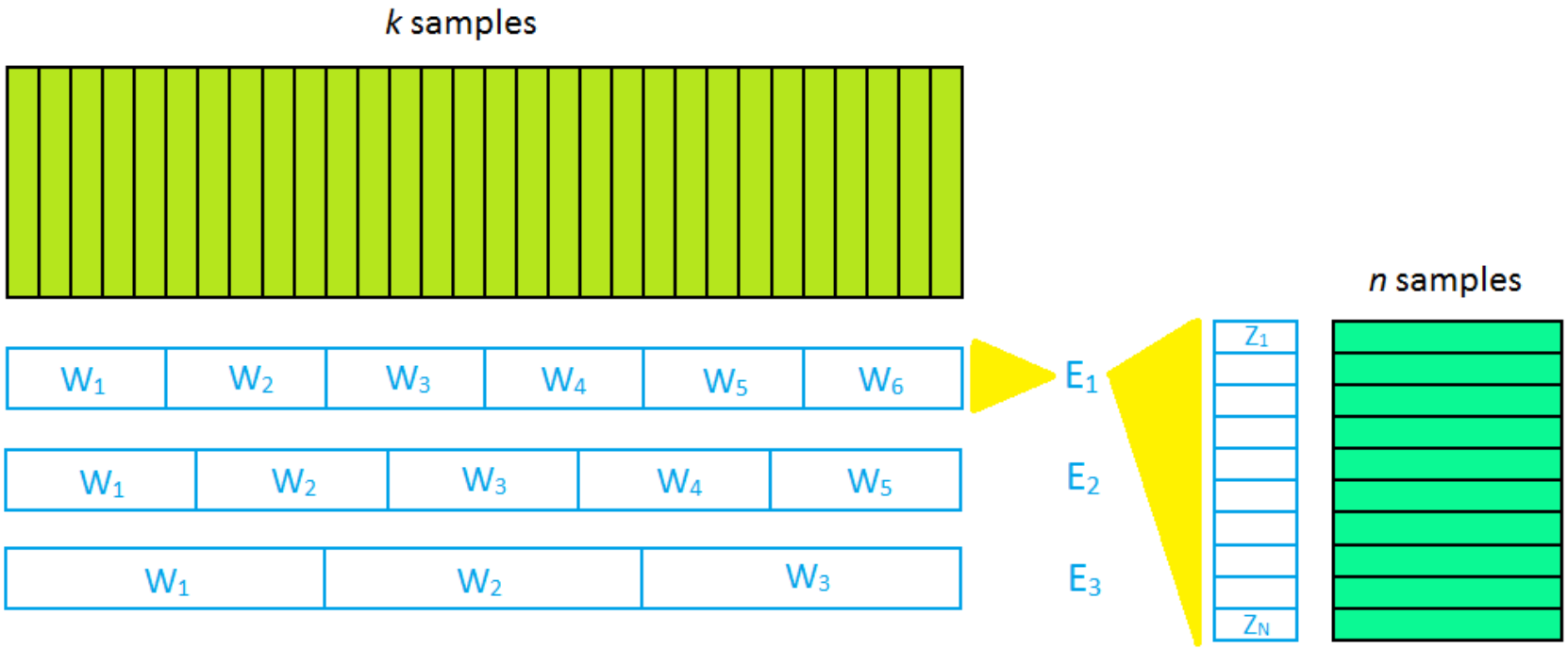




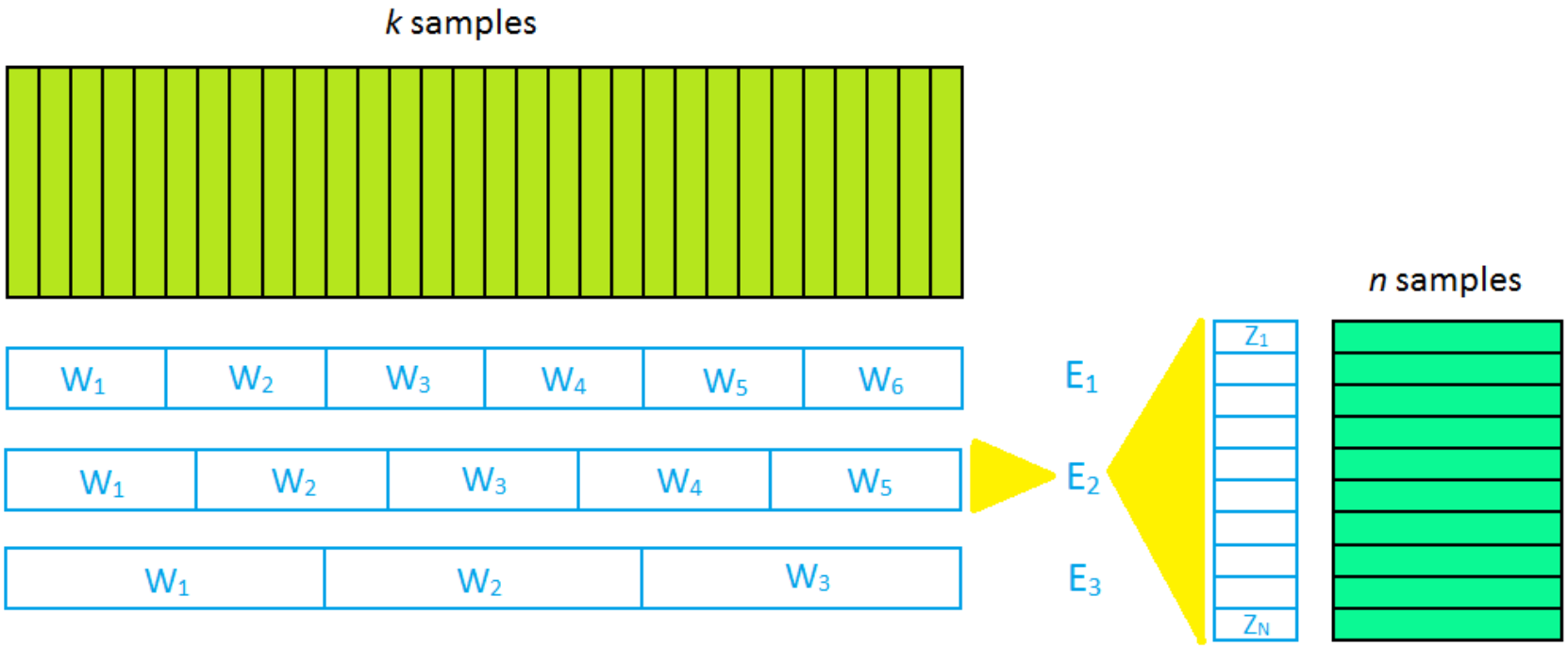
Simulating for multiple *R*'s



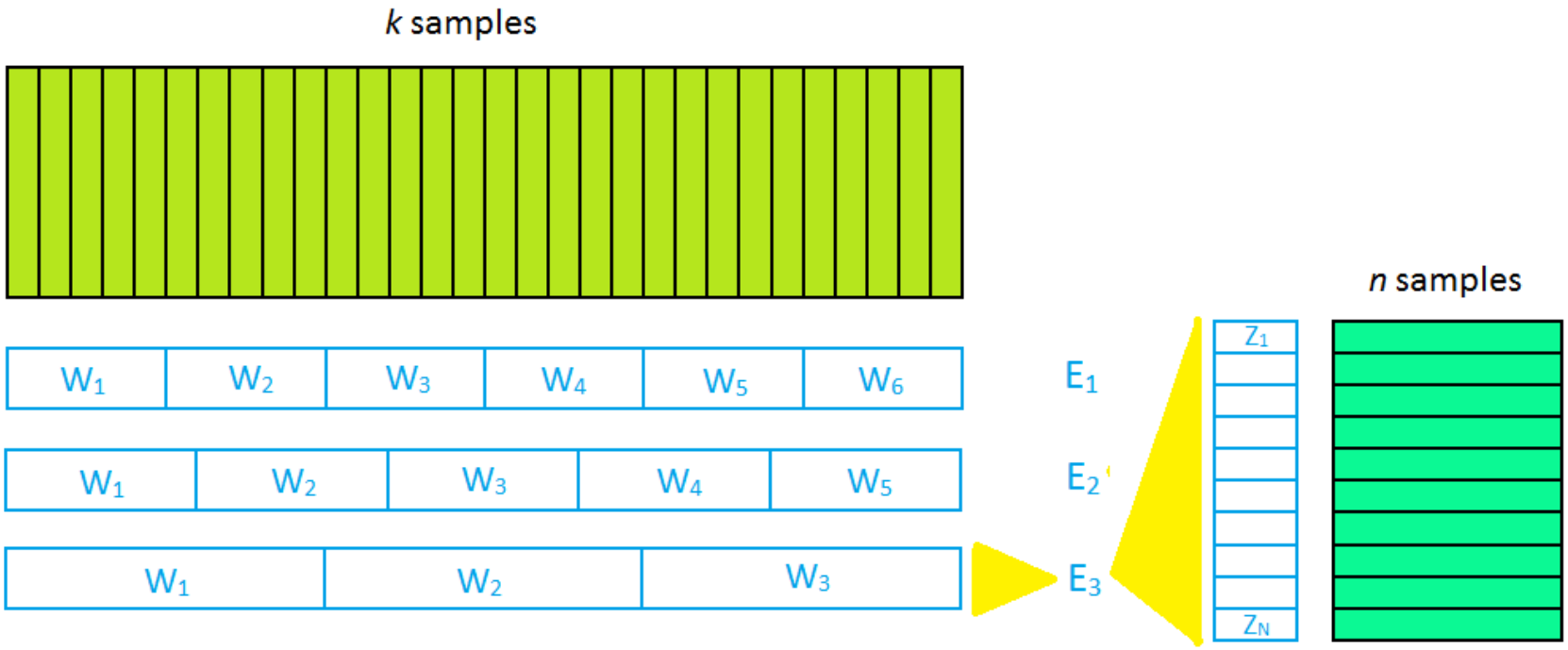
Simulating for multiple *R*'s



Simulating for multiple R's



Simulating for multiple *R*'s



Simulating for multiple R
How to find optimal k and n

- Assume no need for rounding and $m_j = k/R_j, \quad n_j = n$
- Variance for R_j is:

$$\nu_j(m_j, n_j) = \nu_j(k, n) = \frac{R_j}{kn}\hat{G}_{1j} + \frac{1}{n}\hat{G}_{2j} + \frac{R_j}{k}\hat{G}_{3j}$$

- Consider the synthetic criterion function

$$\nu^*(k, n) = \max_{j=1,\ldots,J} \nu_j(k, n)$$

- The function $\nu^*(k, n)$ is convex

Simulating for multiple R

How to find optimal k and n

- Assume time constraint

$$T(k, n) = kt_k + nt_n$$

- Pre-generate k_0 and n_0 samples
- Solve the problem:

$$\begin{cases} \nu^*(k, n) \to min \\ T(k, n) \leq T_0 \\ k \geq k_0 \\ n \geq n_0 \end{cases}$$

  e.g. through Kiefer's (1953) golden section algorithm

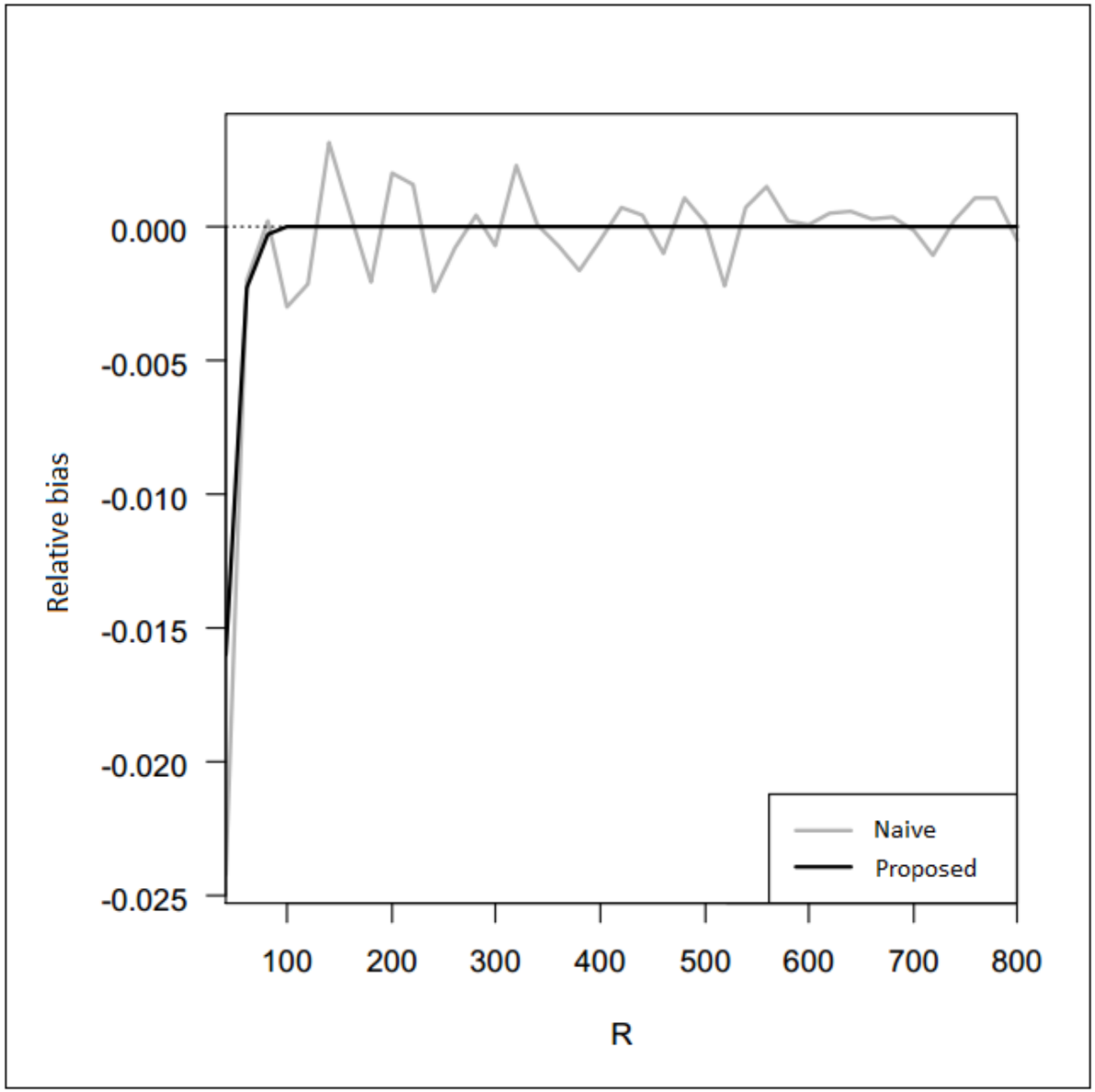
- Draw $k - k_0, \quad n - n_0$ samples
- And proceed with each R_j separately

Example 8
Which procedure better ?

- Population: $\mathbf{c} = [1, 2, ..., 100]'$, $\mathbf{y} = [1, 2, ..., 100]'$
- Budget: $L = 505$ (10% of census cost)
- Estimator $\overline{y}_{EHT}$ with Fattorini's estimates for π's
- Characteristics requested for $R = 40, 60, ..., 800$
- Time limit $T_0 = 7200s$

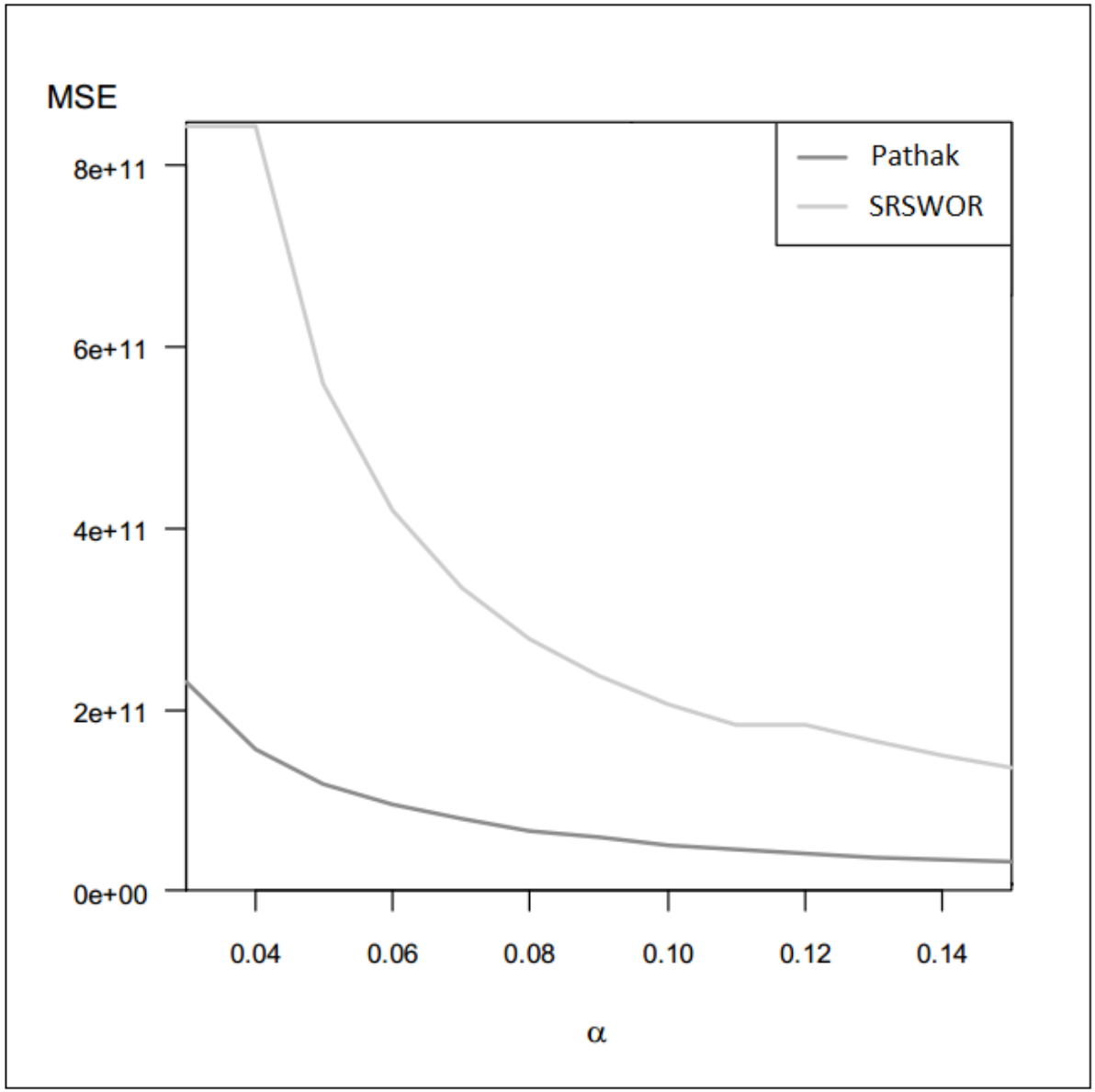
Naive procedure	Proposed procedure
$H = 3000$ full evaluations separately for every R	$k_0 = 6 \cdot 10^3$, $n_0 = 28 \cdot 10^5$ $k = 18008080$, $n = 2907054$
49.257.000 samples time: 7482 s	20.915.134 samples time 7130 s

Example 8
Which procedure better ?



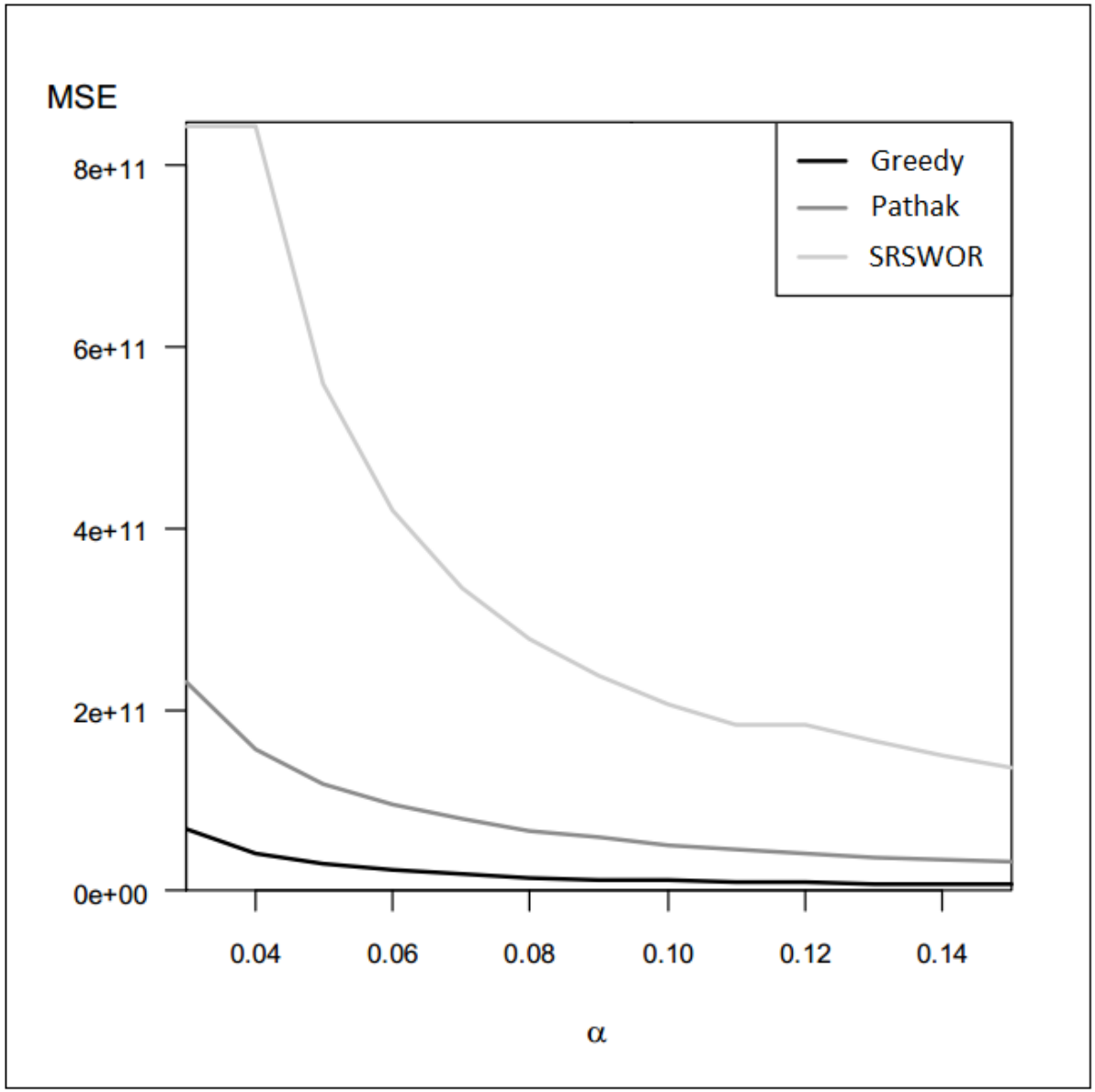
Back *again* to example 5

The graph looked like this



Back *again* to example 5

Now we may compare



Thank you !

Photos: KZK-GOP, Katowice City website, Wikipedia

Another challenge: variance

- For known π's recall the variance of H-T statistic:

$$V(\overline{y}_{HT}) = \frac{1}{N^2} \sum_{i,j \in U} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right)$$

  and its estimator

$$\hat{V}(\overline{y}_{HT}) = \frac{1}{N^2} \sum_{i,j \in s} \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right)$$

- This statistic occasionally becomes negative
- But it is valid for variable-sample-size designs (as opposed to e.g. Yates-Grundy 1953, or Vijayan 1975)

Another challenge: variance

- Fattorini (2006) considered probability estimates

$$\hat{\pi}_{iF} = \frac{k_i + 1}{R + 1} \qquad \hat{\pi}_{ijF} = \frac{k_{ij} + 1}{R + 1}$$

  where k_{ij} is a count of occurences of unit pair (i, j), $i, j \in U$ within R sample replications

- and the statistic:

$$\hat{V}(\overline{y}_{EHT}) = \frac{1}{N^2} \sum_{i,j \in s} \frac{y_i y_j}{\hat{\pi}_{ij}} \left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_i \hat{\pi}_j} - 1 \right) \tag{18}$$

- and has shown its asymptotic unbiasedness for $V(\overline{y}_{EHT})$ when all $\pi_{ij} > 0$

What may go wrong

- Problem 1: usually 2-nd order prob's MUCH lower than 1st! Sampling weights and variance estimates VERY unstable
- Problem 2: Negative variance estimates become common

A modest proposal

- Auxiliary information related to π's often available
- Use it to improve accuracy of probability estimates
- ... and population characteristics estimates

Example 6
Pathak scheme and EHT estimation

Pathak scheme has useful properties (Gamrot 2013) :

Property 1

For $i, i' \in U$ if $c_i \geq c_{i'}$ then $\pi_i \leq \pi_{i'}$

Property 2

For $i \neq j, i' \neq j' \in U$, if $c_i \geq c_{i'}$ and $c_j \geq c_{j'}$ then $\pi_{ij} \leq \pi_{i'j'}$

Property 3

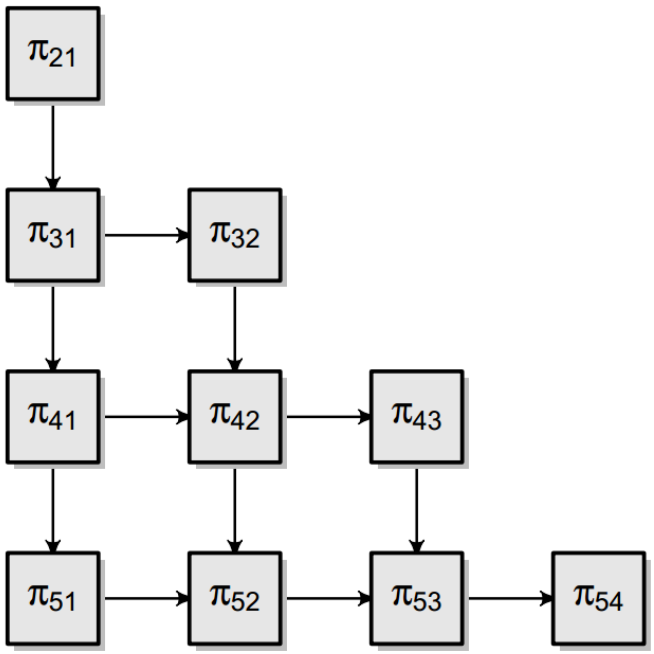
For $i, i' \in U$ if $c_i = c_{i'}$ then $\pi_i = \pi_{i'}$

Property 4

For $i \neq j, i' \neq j' \in U$, if $c_i = c_{i'}$ and $c_j = c_{j'}$ then $\pi_{ij} = \pi_{i'j'}$

Example 6
Pathak scheme and EHT estimation

- Properties 1-4 let us use cost as an auxiliary variable
- Having assumed $c_1 \leq ... \leq c_N$ we obtain
  - $\pi_1 \leq ... \leq \pi_N$
  - $\pi_{i'j'} \leq \pi_{ij}$ for $i \leq i'$ and $j \leq j'$, $i \neq j, i' \neq j' \in U$
- The cover graph for system of inequalities with $N = 5$ is



Example 6
Pathak scheme and EHT estimation

- Fattorini's estimates may violate restrictions
- Forcing restrictions may decrease estimation error:
- Calculation of first-order probabilities:

$$\begin{cases} f_1(\hat{\pi}_1, ..., \hat{\pi}_N) \rightarrow min \\ \hat{\pi}_1 \geq ... \geq \hat{\pi}_N \\ 0 \leq \hat{\pi}_i \leq 1, i \in U \end{cases} \tag{19}$$

where

$$f_1(\hat{\pi}_1, ..., \hat{\pi}_N) = \sum_{i \in U} R_i(\hat{\pi}_{iF} - \hat{\pi}_i)^2 \tag{20}$$

while weights $R_1, ..., R_N$ represent numbers of replications observed for individual units.

Example 6
Pathak scheme and EHT estimation

- Calculation of second-order probabilities:

$$\begin{cases} f_2(\hat{\pi}_{ij}; j < i) \to min \\ \hat{\pi}_{ij} \geq \hat{\pi}_{ij'}; \;\; j, j' < i \in U \\ \hat{\pi}_{ij} \geq \hat{\pi}_{i'j}; \;\; j < i, i' \in U \\ 0 \leq \hat{\pi}_{ij} \leq 1; \;\; j < i \in U \end{cases} \tag{21}$$

where

$$f_2(\hat{\pi}_{ij}, j < i) = \sum_{j<i \in U} R_{ij}(\hat{\pi}_{ijF} - \hat{\pi}_{ij})^2 \tag{22}$$

while R_{ij} for $i > j \in U$ represent numbers of replications containing both i-th and j-th unit.

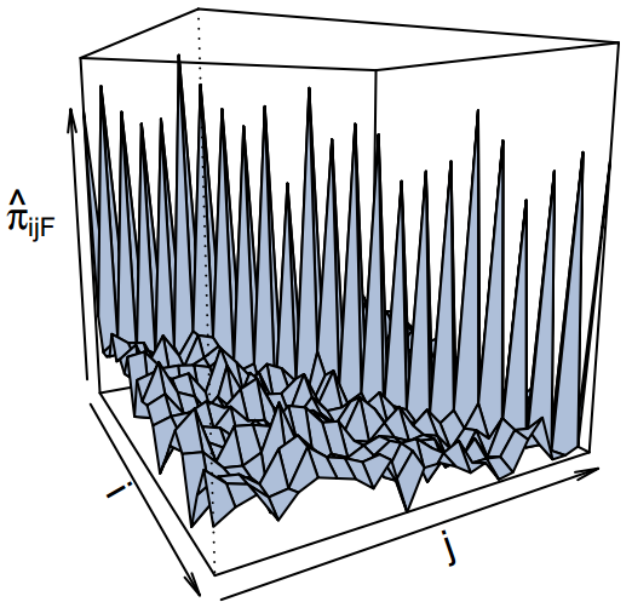
- Both solved by Active Set Method (deLeeuw et al 2009)

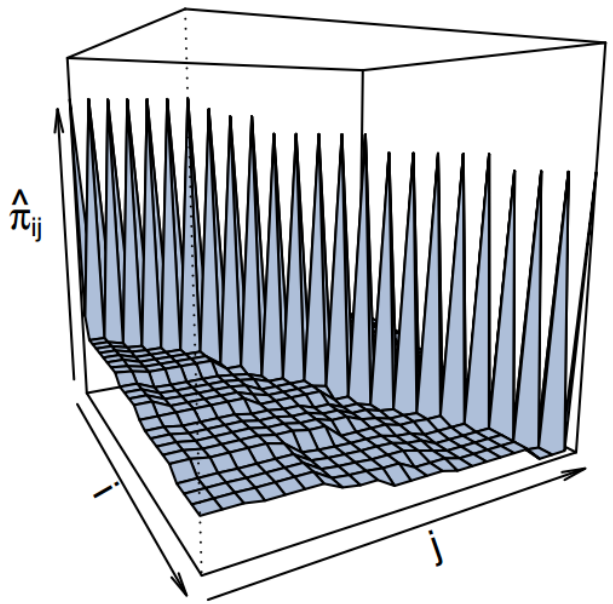
Example 6 - Illustration
Pathak scheme and EHT estimation

- Population of $N = 24$ units
- Individual costs $\mathbf{c} = [1, ..., 24]'$
- Survey budget $L = 90$ (30% of census cost)
- $R = 500$ sample replications

Example 6 - Illustration
Pathak scheme and EHT estimation

Estimated matrices of inclusion probabilities





Example 6 - Simulation

Pathak scheme and EHT estimation

Process repeated 5000 times for various **y**-vectors

y	P_F	P_A	P_A/P_F,	V_A/V_F	M_A/M_F
$[1, 2, ..., 24]'$	0.3695	0.2134	0.5776	0.4329	0.5028
$[1, ..., 12, 12, ..., 1]'$	0.0447	0.0026	0.0581	0.6647	0.7269
$[12, ..., 1, 1, ..., 12]'$	0.1365	0.0544	0.3985	0.7076	0.7537
$[24, 23, ..., 1]'$	0.0054	0.0003	0.0645	0.9853	1.0008

P_F, P_A - percent of negative estimates (for Fattorini / Active Set)
V_F, V_A - variance of variance estimates
M_F, M_A - MSE of variance estimates

Example 6 - Simulation
Pathak scheme and EHT estimation

Remarks:

- Large values of P_F, P_A due to small R
- Forcing restrictions tends to improve estimates
- Simulations results for whole population indirectly utilized
- Eventual ties shoud lead to further improvement
- Feasibility restricted by the size of the problem

References

- Aires N. (1999) Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. *Methodology and Computing in Applied Probability*, 4, 457-469.
- Bakhshi Z. H., Khan M.F., Ahmad Q. S. (2010), Optimal Sample Numbers in Multivariate Stratified Sampling with a Probabilistic Cost Constraint, *Int J. of Math and Appl Stat* 1(2), 111-120.
- Breidt F.J. Fuller W.A. (1993) Regression Weighting for Multiphase Samples, *Sankhya* Special Volume 55, Series B, 297-309
- Javed S., Bakhshi Z.H., Khalid M.M. (2009) Optimum allocation in Stratified Sampling with random costs, *Int Rev Pure Appl Math*. 5(2), 363-370.
- Kadane, J.B. (2005). Optimal Dynamic Sample Allocation Among Strata. *J. Official Stat*, 21, 531-541
- Land A.H., Doig A.G. (1960) An automatic method of solving discrete programming problems, *Econometrica* 28(3), 497-520.
- Khan M.F., Ali I., Raghav Y.S., Bari A. (2012) Allocaton in Multivariate Stratified Surveys with Non-Linear Random Cost Function, *Am. J. of Oper Res*, 2, 100-105.
- Rosen B. (1997) Asymptotic theory of order sampling *J. Stat. Plann. Inf.* 62, 135-158.

References

- Fattorini L. (2006) Applying the Horvitz-Thompson Criterion in Complex De-signs: A Computer-Intensive Perspective for Estimating Inclusion Probabilities,*Biometrika*, 93, (2), 269-278.
- Fattorini L. (2009) An adaptive algorithm for estimating inclusion probabilities and performing the Horvitz-Thompson criterion in complex designs. *Comput. Stat.* 24, 623 639.
- Pathak K. (1976) Unbiased Estimation in Fixed-Cost Sequential Sampling Schemes, *Ann. Stat.*, 4 (5), 1012-1017.
- Thompson M. E. Wu C. (2008) Simulation-based randomized systematic PPS sampling under substitution of units, *Surv. Meth.*, 34 (1), 3-10.
- Gamrot W. (2012) Simulation-assisted Horvitz-Thompson statistic and isotonic regression. In: Ramik J., Stavarek D. (eds.) *Proceedings of the 30th MME Conference*, 207-212.
- Gamrot W. (2013) On exact computation of minimum sample size for restricted estimation of binomial parameter, *J. Stat. Plan. Inf.* 143, 852-866.
- Gamrot W. (2013) On Kernel Smoothing and Horvitz-Thompson Estimation. *Studia Ekonomiczne*. 152, 32-41.
- Gamrot W. (2014) Estimators for the Horvitz-Thompson Statistic Based on Some Posterior Distributions. *Math Popul Stud*, 21(1), 12-29.
- Gamrot W. (2015) *Estymacja wartości przeciętnej uwzględniająca koszt pozyskania danych*. UE. Katowice.