Adaptive Probability-Based Latin Hypercube Designs for Slid-Rectangular Regions

Ying Hung
Department of Statistics, Rutgers University

Overview

- Introduce space-filling designs
- A new class of space-filling designs for a specific type of irregular regions
- Idea of adaptive designs
- Unbiased estimators
- Illustrations

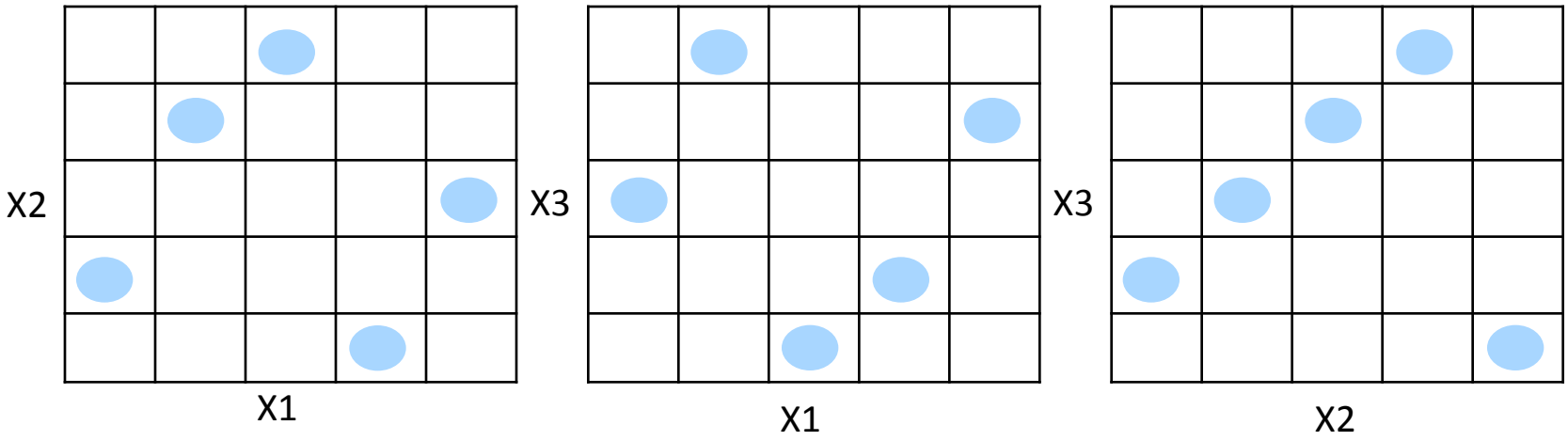
Latin Hypercube Designs (LHDs)

- A n-run LHD can be generated using a random permutation of $\{1, 2, \ldots, n\}$ for each factor.

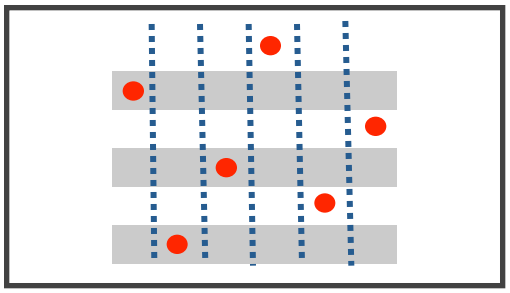
- Space-filling design: One-dimensional balance.

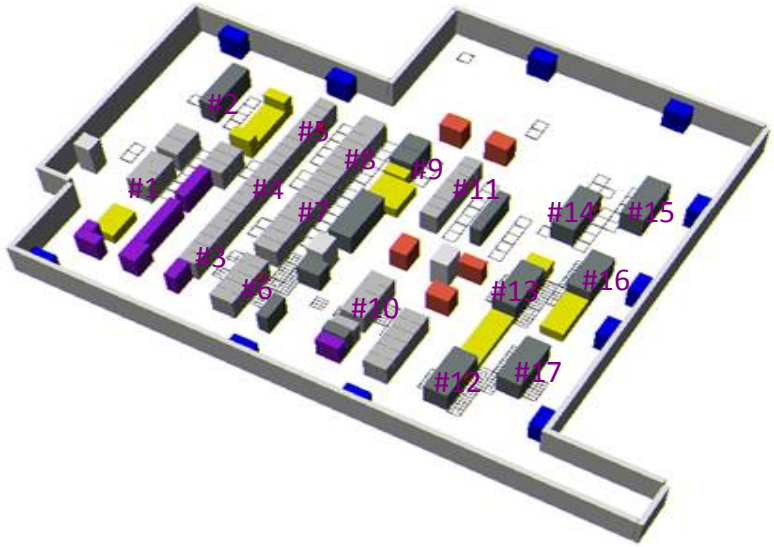
- Limitation: Constructed based on rectangular regions.

run	X1	X2	X3
1	1	2	3
2	2	4	5
3	3	5	1
4	4	1	2
5	5	3	4



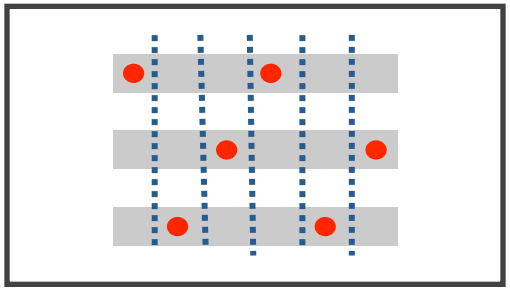
A Naïve Strategy with Irregular Regions

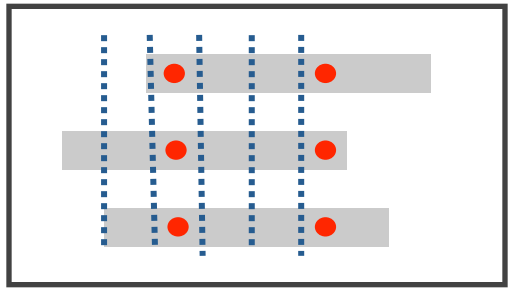




collapse

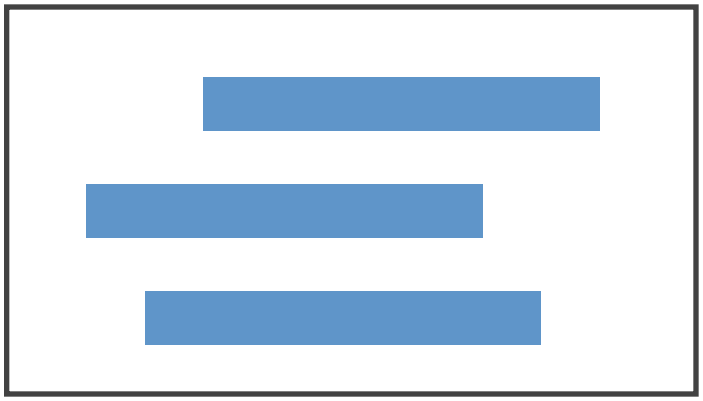
But … Irregular experimental region





Q: How to construct space-filling design with irregular experimental region?

Probability-Based Latin Hypercube Designs (PLHD)

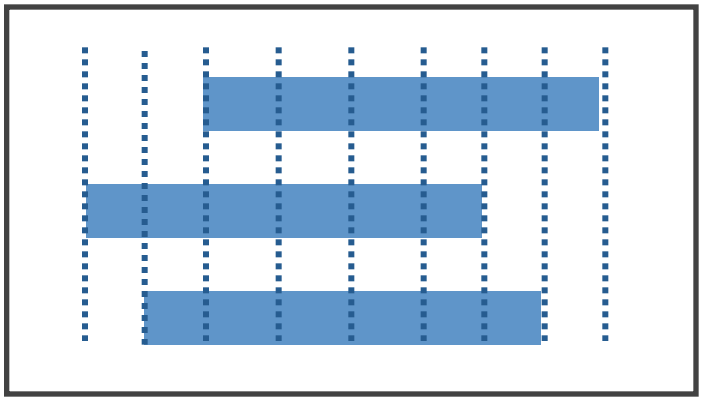


Slid-rectangular regions:
a specific type of irregular regions, where the desirable range of one factor depends on the level of another factor.

How to construct design with the following properties?

- *One-dimensional balance*
- *Number of design points is proportional to the length of experimental region*

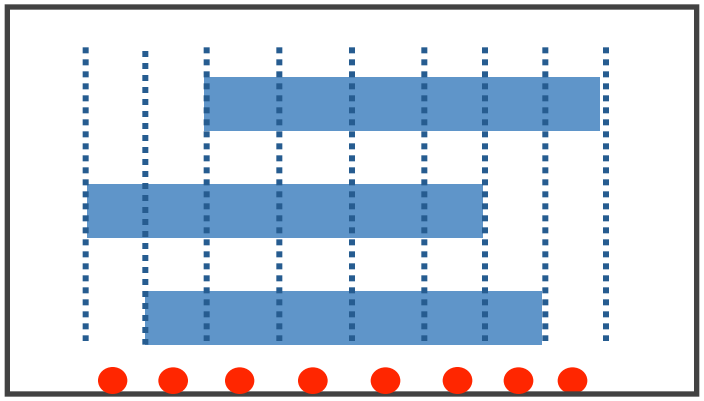
Probability-Based Latin Hypercube Designs (PLHD)



How to construct design with the following properties?

- *One-dimensional balance*
- *Number of design points is proportional to the length of experimental region*

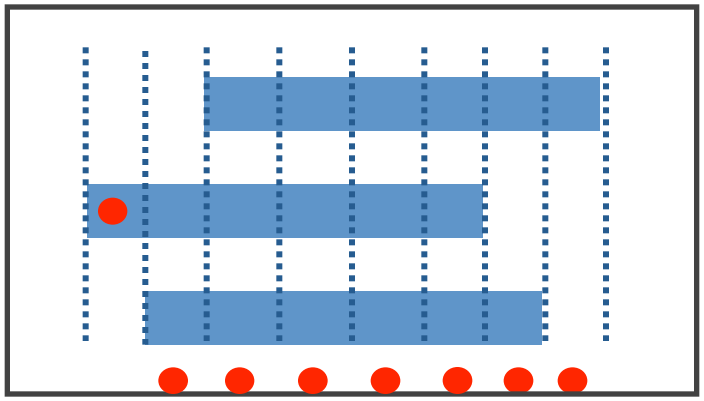
Probability-Based Latin Hypercube Designs (PLHD)



How to construct design with the following properties?

- *One-dimensional balance*
- *Number of design points is proportional to the length of experimental region*

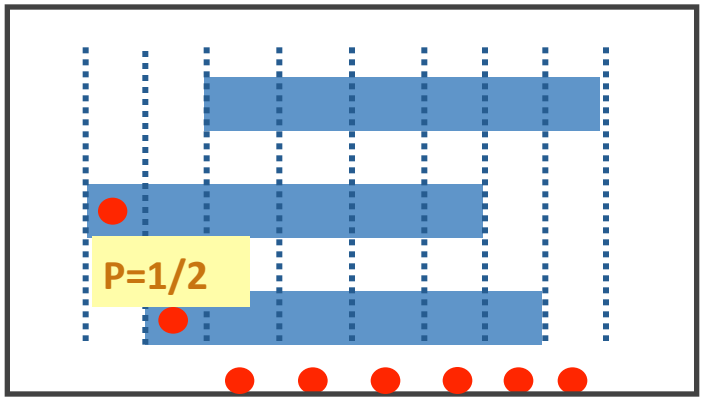
Probability-Based Latin Hypercube Designs (PLHD)



How to construct design with the following properties?

- *One-dimensional balance*
- *Number of design points is proportional to the length of experimental region*

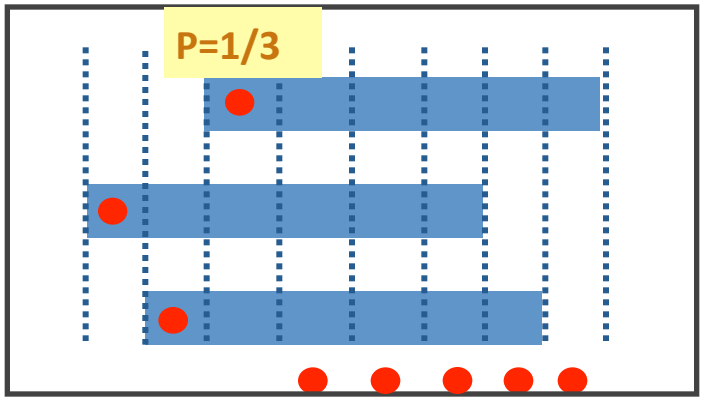
Probability-Based Latin Hypercube Designs (PLHD)



How to construct design with the following properties?

- *One-dimensional balance*
- *Number of design points is proportional to the length of experimental region*

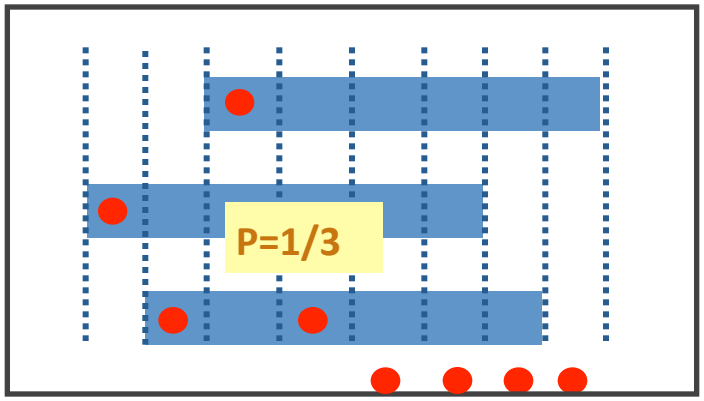
Probability-Based Latin Hypercube Designs (PLHD)



How to construct design with the following properties?

- *One-dimensional balance*
- *Number of design points is proportional to the length of experimental region*

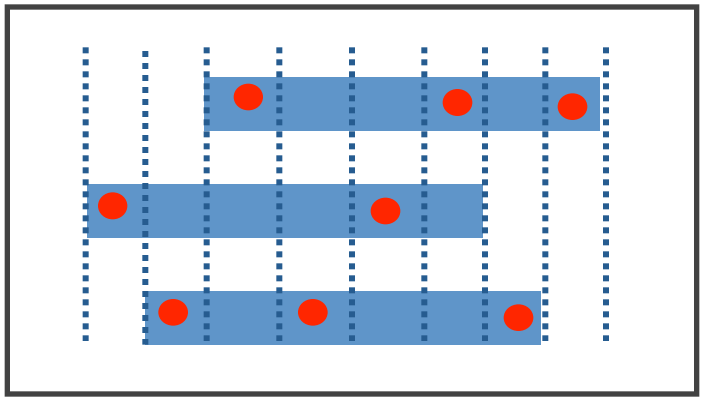
Probability-Based Latin Hypercube Designs (PLHD)



How to construct design with the following properties?

- *One-dimensional balance*
- *Number of design points is proportional to the length of experimental region*

Probability-Based Latin Hypercube Designs (PLHD)



Optimal Design Criteria

- Maximin distance
- Minimize correlation
- etc.

New Algorithm

- A new heuristic algorithm for optimal PLHD searching.

How to construct design with the following properties?

- *One-dimensional balance*
- ***Number of design points is proportional to the length of experimental region***

Reference: Y. Hung, Y. Amemiya, and C. F. Jeff Wu (2010). Probability-Based Latin Hypercube Design, *Biometrika*, 97, 961-968.

Definition of Probability-Based Latin Hypercube Designs

- Assume there are p factors and the first two factors, x_1 and x_2, form a slid-rectangular region.
- Factor x_2 has k levels and the ranges for x_1 are located irregularly on a interval $[A, B]$. For the jth level of x_2, the feasible interval for x_1 is denoted by (A_j, B_j).
- $A=\min\{A_j\}$, $B=\max\{B_j\}$
- Divide the interval $[A, B]$ into n equally spaced sub-intervals and assigning the n levels of by the middle ofx_1 these subintervals.
- For each level of x_1, the feasible range of x_2 is defined by C_i, and the level of x_2 is assigned by

$$\mathrm{pr}(x_{2i} = j) = \begin{cases} [\sum_{j=1}^{k} I(j \in C_i)]^{-1}, & \text{if } j \in C_i, \\ 0, & \text{otherwise.} \end{cases}$$

Unbiased Estimator

- An unbiased estimator of the population mean based on the probability-based LHDs can be written as

$$T = N^{-1} \sum_{i=1}^{n} \sum_{j \in C_i} E(w_{ij})^{-1} w_{ij} g(Y_{ij}),$$

where $g(.)$ is an arbitrary function, Y_{ij}s are the responses, w_{ij} is an indicator variable with w_{ij}=1 if Y_{ij} is selected by the design, and w_{ij}=0 otherwise.

- Its variance can be written in the Yates-Grundy expression (Cochran, 1977, pp. 260)

$$\text{var}(T) = \tfrac{1}{2} N^{-2} \left\{ \sum_{ij} \sum_{qt(qt \neq ij)} (\pi_{ij}\pi_{qt} - \pi_{ij,qt}) [\frac{g(Y_{ij})}{\pi_{ij}} - \frac{g(Y_{qt})}{\pi_{qt}}]^2 \right\},$$

$\pi_{ij} = E(w_{ij}) = c_i^{-1}$, $\pi_{ij,it} = 0$ for $t \neq j$, and $\pi_{ij,qt} = \pi_{ij}\pi_{qt} = c_i^{-1} c_q^{-1}$ for $i \neq q$.

Further Improvement

- PROPOSITION 1. *Let n_j denote the number of points with $x_2 = j$. Then*

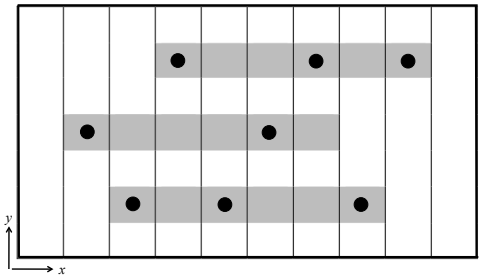
$$E(n_j) = \sum_{i=1}^{n} \mathrm{pr}(x_{2i} = j) = \sum_{i=1}^{n} \frac{I(j \in C_i)}{\sum_{j=1}^{k} I(j \in C_i)}.$$

- Proposition 1 shows the expected number of points located in each shaded area. For example:

  17/6 (= 4/3+1/2+1) for the upper shaded area,

  17/6 (= 1+1/2+4/3) for the middle one,

  7/3 (= 1/2+4/3+1/2) for the lower one.



- Ideally, the n_j values in Proposition 1 should be proportional to the length of the shaded area, which can be written as B_j-A_j. This is because the information from each area is assumed to be proportional to its length.

Balanced PLHD

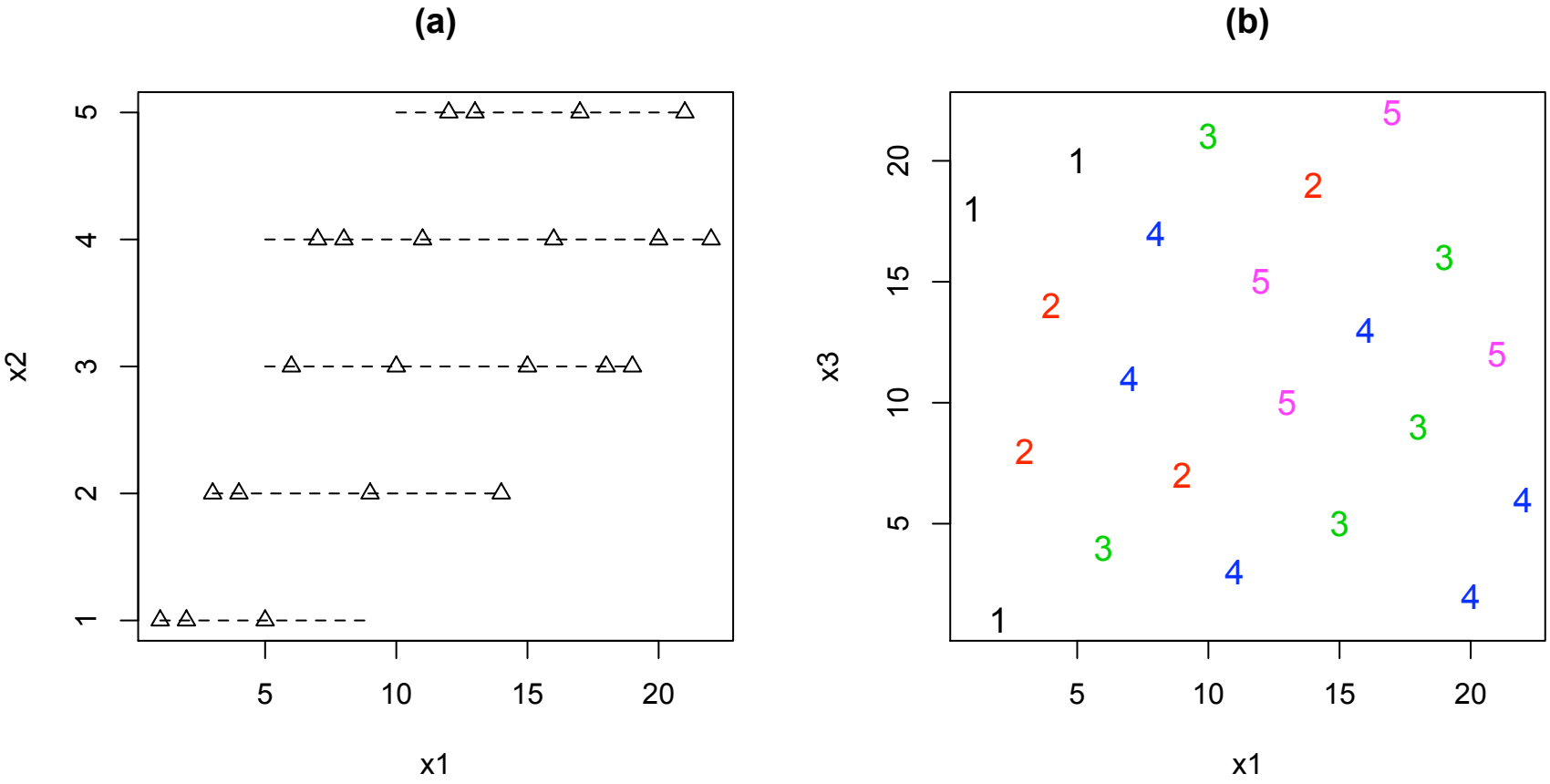
- Inspired by the observation in Proposition 1, a modification is introduced to incorporate the proportional balance property, where the number of observations is proportional to the length of the interval.
- We call this design a balanced probability-based LHD. It can be written as modified probability-based LHDs with the constraints

$$\frac{n_j}{n} = \frac{(B_j - A_j)}{\sum_{j=1}^{k}(B_j - A_j)} = p_j, \text{ for } j = 1, \ldots, k.$$

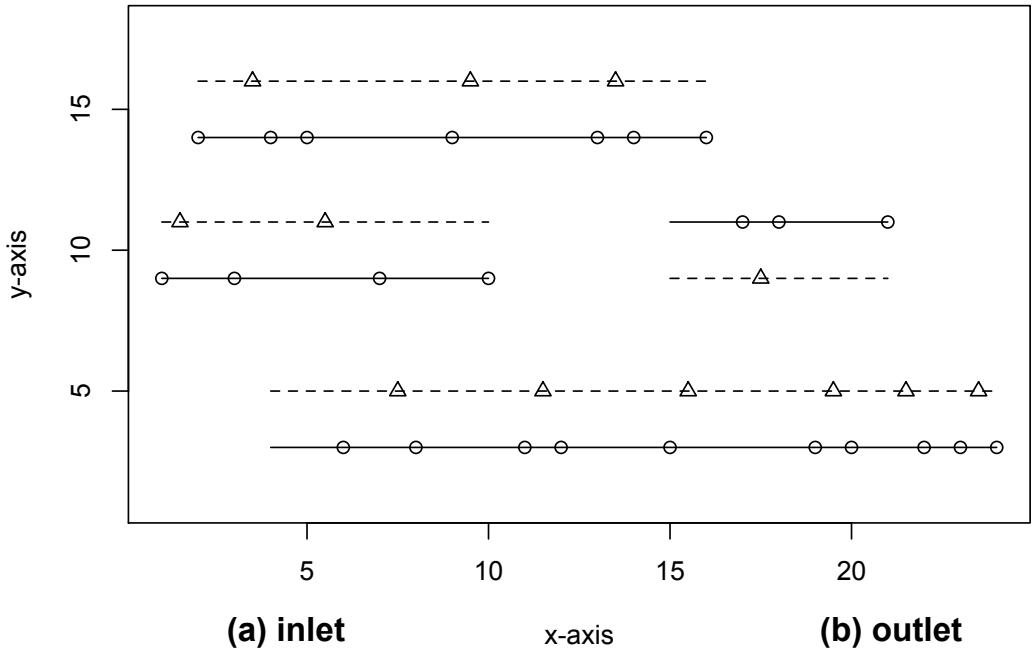
- In practice, the quantities np_j are not always integers for given n and p_j. In that situation, an approximate balance with $|n_j\text{-}np_j| < 1$ should be imposed.

Example of Balanced PLHD

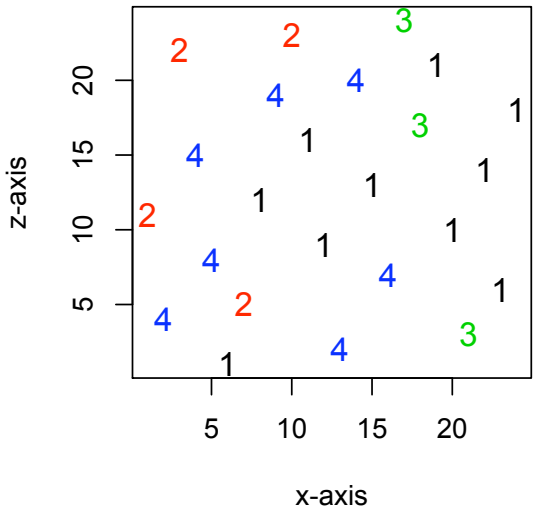
- A design with three factors and 22 runs. For the slid-rectangular region, factor x_2 has five levels and the proportional lengths of x_1 at different levels of x_2 are 3:4:5:6:4.

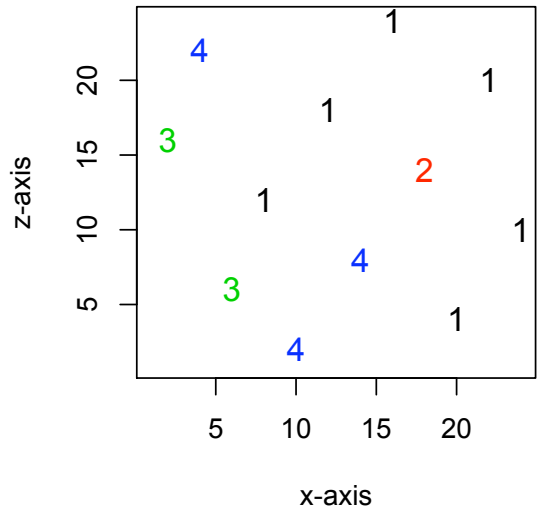


A Data Center Example



(a) inlet **(b) outlet**

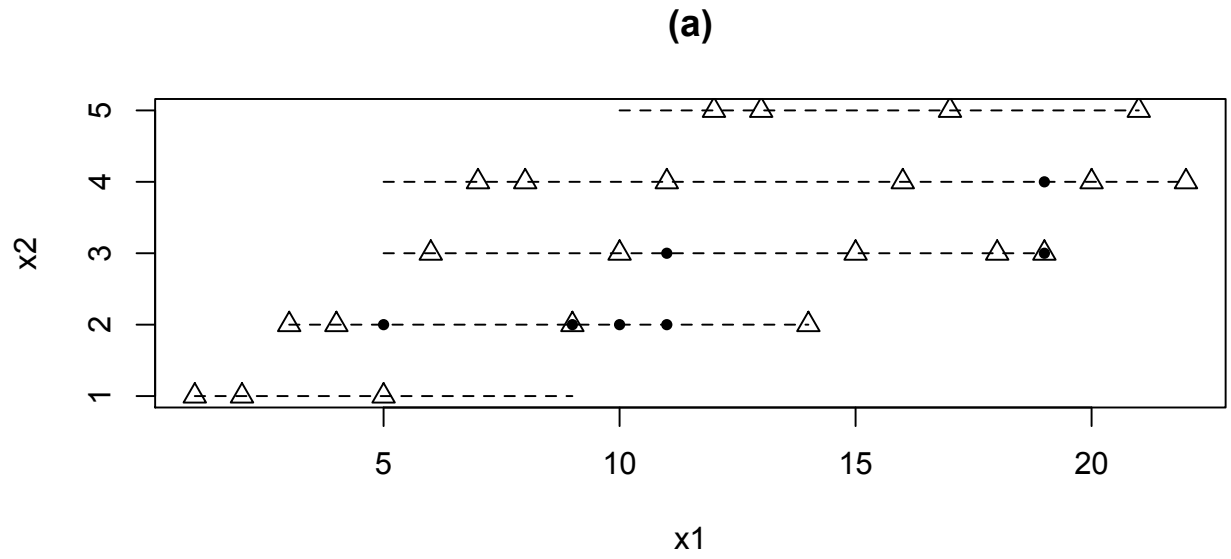


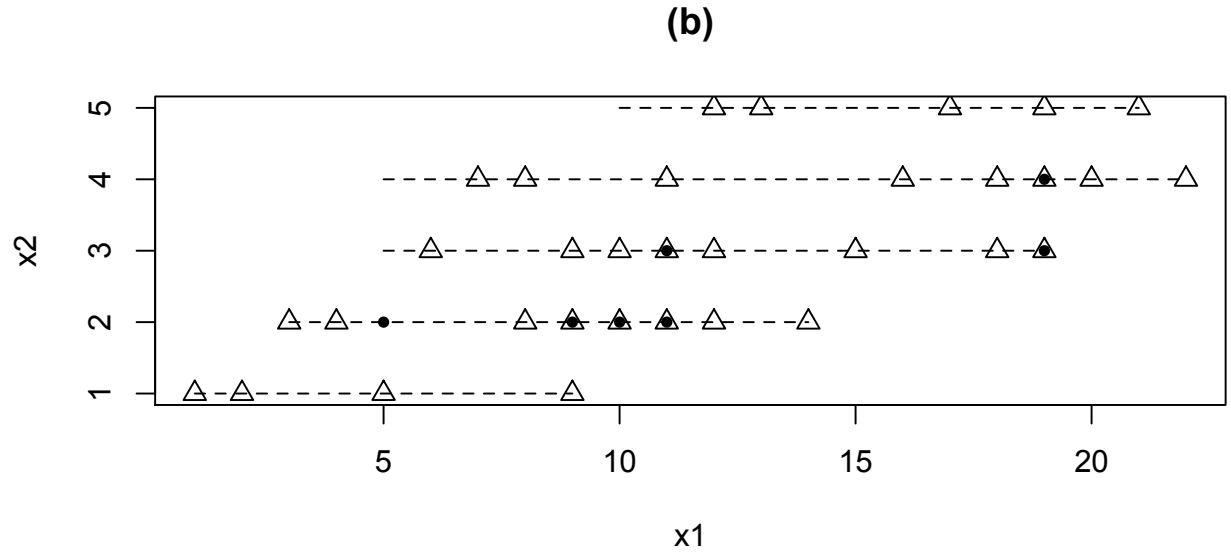


Next Question…

- How to construct a plan that can adaptively increase sample effort in the neighborhood of the high temperature observations.
- Idea:
  - Construct a probability-based LHD to collect initial sample
  - Based on *n* initial observations, whenever the response of a selected unit satisfies a given criterion, additional units in the neighborhood of that unit are added to the sample.
  - This procedure continues until no more units are found that meet the condition.
  - The final design contains every unit in the neighborhood of any sample unit satisfying the condition.

Adaptive Probability-based LHDs





Definition

- Network: The set of all units satisfying the condition in the neighborhood of one another. Section in the initial design of any point in a network will result in the final sample of all units in that sample.

- For any unit that does not meet the condition, it forms a network with size one.

- The population can be partitioned into K networks.

- The units that do not meet the condition but in the neighborhood of some design points that satisfy the condition are call edge points.

Reference: Ying Hung (2011). Adaptive Probability-based Latin Hypercube Designs, *the Journal of American Statistical Association*, 106, 213-219.

Unbiased Estimators for Adaptive Designs

- Let Ψ_k be the set of units in the kth network. The number of units selected from the kth network in the initial sample is

$$n_k = \sum_{i \in \Psi_k} I(i \in s_0),$$

and the inclusion probability of network k is

$$P(n_k > 0) = 1 - \frac{\prod_{l \in \psi_k} (c_l - d_{kl})}{\prod_{l \in \psi_k} c_l}.$$

- A unbiased estimator for adaptive probability-based PLHDs

$$\hat{\mu} = \frac{1}{N} \sum_{k}^{K} \frac{y_k^* I(n_k > 0)}{P(n_k > 0)},$$

where K is the number of networks in the population, $I(n_k>0)$ takes 1 if any unit of the kth network is in the initial sample and $y_k^* = \sum_{j \in \Psi_k} y_j$.

Variance of the Unbiased Estimator

The variance of the unbiased estimator can be calcuated by

$$var(\hat{\mu}) = N^{-2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* [P(n_k > 0, n_h > 0) - p(n_k > 0) P(n_h > 0)]}{P(n_k > 0) P(n_h > 0)},$$

where

$$P(n_k > 0, n_h > 0) = 1 - \frac{\prod_{l \in \psi_k} (c_l - d_{kl})}{\prod_{l \in \psi_k} c_l} - \frac{\prod_{l \in \psi_h} (c_l - d_{hl})}{\prod_{l \in \psi_k} c_l} + \frac{\prod_{l \in (\psi_k \cup \psi_h)} (c_l - d_{k \cup h, l})}{\prod_{l \in (\psi_k \cup \psi_h)} c_l}.$$

and $d_{k \cup h, l}$ is the number of units in $\Psi_{kl} \cup \Psi_{hl}$. An unbiased estimator of the variance of $\hat{\mu}$ is

$$\hat{var}(\hat{\mu}) = N^{-2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k^* y_h^* [P(n_k > 0, n_h > 0) - p(n_k > 0) P(n_h > 0)]}{P(n_k > 0) P(n_h > 0) P(n_k > 0, n_h > 0)} I(n_k > 0) I(n_h > 0).$$

Improved Unbiased Estimator

- The foregoing unbiased estimator can be improved by incorporating more of the information in the final sample.
- For example, the observations from edge points are used in the estimator only they are included in the initial sample.
- Using the Rao-Blackwell method, an improved unbiased estimator can be obtained by calculating the conditional expectation of the original estimator, given a sufficient statistics.
- The most efficient choice is the minimum sufficient statistics.

Rao-Blackwell Unbiased Estimator

- The minimum sufficient statistics m is the unordered set of distinct labeled observations, i.e., $m = \{(i, y_i) : i \in s\}$.
- Define M as the sample space of m, $g(s_0)$ as the function that maps an initial design s_0 into a value of m, and S as the sample space containing all possible samples.
- The resulting unbiased estimator

$$\begin{aligned}\hat{\mu}^{\mathrm{RB}} &= E(\mu|M=m) \\ &= \frac{1}{N}\sum_{k=1}^{K}\frac{y_k^* J_k(1-e_k^*)}{P(n_k>0)} + \frac{1}{NL}\sum_{s_0'\in S}\left\{I(g(s_0')=m)\left[\sum_{i\in s_0, e_i=1} y_i c_i\right]\right\},\end{aligned}$$

where $e^* = \sum_{i\in\Psi_k} e_i$ and $e_i = 1$ if unit i is an edge point and $e_i = 0$ otherwise.

- The variance can be written as

$$var(\hat{\mu}^{\mathrm{RB}}) = var(\hat{\mu}) - \sum_{m \in M} \frac{P(m)}{L} \sum_{\tilde{s}_0 \in S} I(g(\tilde{s}_0) = m)\left[(\hat{\mu} - \hat{\mu}^{\mathrm{RB}})^2\right],$$

where L is the number of initial designs that are *compatible* with m and $P(m)$ is the probability that $M = m$. An unbiased estimator of the variance is

$$\tilde{var}(\hat{\mu}^{\mathrm{RB}}) = \hat{var}(\hat{\mu}) - L^{-1} \sum_{\tilde{s}_0 \in S} I(g(\tilde{s}_0) = m)\left[(\hat{\mu} - \hat{\mu}^{\mathrm{RB}})^2\right]$$

and a more efficient estimator can be further obtained by conditioning on the minimum sufficient statistics as follows

$$\begin{aligned} \hat{var}(\hat{\mu}^{\mathrm{RB}}) = & \; E(\tilde{var}(\hat{\mu}^{\mathrm{RB}}) \mid M = m) \\ = & \; \tfrac{1}{L} \textstyle\sum_{\tilde{s}_0 = S} I(g(\tilde{s}_0) = m) \hat{var}(\hat{\mu}) - \tfrac{1}{L} \textstyle\sum_{\tilde{s}_0 \in S} I(g(\tilde{s}_0) = m)\left[(\hat{\mu} - \hat{\mu}^{\mathrm{RB}})^2\right]. \end{aligned}$$

Further Improvement

- Although conditioning on the minimum sufficient statistics is the most efficient one, it is computationally difficult for large designs because one has to evaluate all the compatible designs in order to obtain the estimation.

- Idea: Construct an unbiased estimator by conditioning on a carefully chosen sufficient statistics, instead of the minimum sufficient statistics.

Further Improved Unbiased Estimator

Let s denote the final sample and define s_c as the set of all the distinct units in the sample for which the condition to sample adaptively is satisfied. The remaining part is denoted by $s_{\bar{c}}$. Define V as a collection of x_1 coordinates with which edge points occurs in the initial sample. For unit i, let f_i be the number of times that the network to which unit i belongs is intersected by the initial sample. Using the above notation, a sufficient statistics can be defined by

$$m^* = \{(i, y_i, f_i), V, (j, y_j) : i \in s_c, j \in s_{\bar{c}}\},$$

and the sample space for m^* is defined by M^*. Hence, the improved unbiased estimator by conditioning on the sufficient statistics m^* can be obtained by

$$\begin{aligned} \hat{\mu}^* = & \quad E(\mu | M^* = m^*) \\ = & \quad \frac{1}{N} \sum_{k=1}^{K} \frac{y_k^* I(n_k > 0)\left(1 - e_k^*\right)}{P(n_k > 0)} + \frac{1}{N} \sum_{l \in V} \frac{\sum_{i \in s} e_i y_i t_l(i)}{e_{s_l} c_l^{-1}}, \end{aligned}$$

where $t_l(i)$ is an indicator variable taking 1 if the unit i belongs to level l in factor x_1 (column l) and 0 otherwise and $e_{s_l} = \sum_{i \in s} e_i t_l(i)$.

The variance of the improved unbiased estimator is

$$var(\hat{\mu}^*) = var(\hat{\mu}) - \sum_{m^* \in M^*} \frac{P(m^*)}{L} \sum_{s_0' \in S} \left\{ I(g(s_0') = m^*) \left[\sum_{l=1}^{n} \frac{c_l}{N} \left(\sum_{i \in s_0', e_i = 1} y_i t_l(i) - \frac{1}{e_{s_l}} \sum_{i \in s} e_i y_i t_l(i) \right) \right]^2 \right\},$$

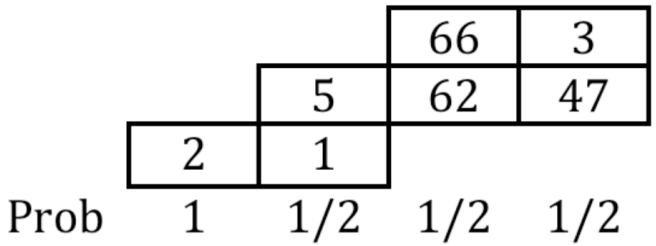
An unbiased estimator of the variance is

$$\tilde{var}(\hat{\mu}^*) = \hat{var}(\hat{\mu}) - \frac{1}{L} \sum_{s_0' \in S} \left\{ I(g(s_0') = m^*) \left[\sum_{l=1}^{n} \frac{c_l}{N} \left(\sum_{i \in s_0', e_i = 1} y_i t_l(i) - \frac{1}{e_{s_l}} \sum_{i \in s} e_i y_i t_l(i) \right) \right]^2 \right\}$$

and a more efficient estimator of the variance can be obtained by

$$\begin{aligned} \hat{var}(\hat{\mu}^*) &= E[\tilde{var}(\hat{\mu}^*) \mid M^* = m^*] \\ &= \frac{1}{L} \sum_{s_0' \in S} I(g(s_0') = m^*) \hat{var}(\hat{\mu}(s_0')) - \frac{1}{L} \sum_{s_0' \in S} \left\{ I(g(s_0') = m^*) \left[\sum_{l=1}^{n} \frac{c_l}{N} \left(\sum_{i \in s_0', e_i = 1} y_i t_l(i) - \frac{1}{e_{s_l}} \sum_{i \in s} e_i y_i t_l(i) \right) \right]^2 \right\}. \end{aligned}$$

Simulation 1: Probability-based LHD



- All possible adaptive PLHDs and a comparison of the unbiased estimators

sampler	V	$\hat{\mu}$	$\hat{\mu}^*$	$\hat{\mu}^{\mathrm{RB}}$	$\hat{\mathrm{var}}(\hat{\mu})$	$\hat{\mathrm{var}}(\hat{\mu}^*)$	$\hat{\mathrm{var}}(\hat{\mu}^{\mathrm{RB}})$
$\{11\},\{21\},\{32\},\{42\};\{33\},\{22\},\{43\}$	4	32.29	26	26	591.76	277.47	320.24
$\{11\},\{21\},\{32\},\{43\};\{33\},\{22\},\{42\}$	4	19.71	26	26	42.20	277.47	320.24
$\{11\},\{21\},\{33\},\{42\};\{32\},\{42\},\{22\}$	4	32.29	26	26	591.76	277.47	320.24
$\{11\},\{21\},\{33\},\{43\};\{32\},\{22\},\{42\}$	4	19.71	26	26	42.20	277.47	320.24
$\{11\},\{22\},\{32\},\{42\};\{33\},\{43\}$	2, 4	33.43	27.14	27.14	634.53	320.24	320.24
$\{11\},\{22\},\{32\},\{43\};\{33\},\{42\}$	2, 4	20.86	27.14	27.14	84.98	320.24	320.24
$\{11\},\{22\},\{33\},\{42\};\{32\},\{43\}$	2, 4	33.43	27.14	27.14	634.53	320.24	320.24
$\{11\},\{22\},\{33\},\{43\};\{32\},\{42\}$	2, 4	20.86	27.14	27.14	84.98	320.24	320.24
Mean		26.57	26.57	26.57	338.37	298.86	298.86

Illustration

$$\hat{\mu} = \frac{1}{7}\left[\frac{y_{11}}{1} + \frac{y_{21}}{1/2} + \frac{(y_{32}+y_{33})}{1} + \frac{y_{42}}{1/2}\right]$$
$$= \frac{1}{7}\left[\frac{2}{1} + \frac{1}{1/2} + \frac{(62+66)}{1} + \frac{47}{1/2}\right] = 32.29.$$

$$\hat{\mu}^{\mathrm{RB}} = \frac{1}{7}\left[\frac{y_{11}}{1} + \frac{y_{21}}{1/2} + \frac{(y_{32}+y_{33})}{1} + \frac{2y_{42}+2y_{43}+2y_{42}+2y_{43}}{4}\right]$$
$$= \frac{1}{7}\left[\frac{2}{1} + \frac{1}{1/2} + \frac{(62+66)}{1} + \frac{94+6+94+6}{4}\right] = 26.$$

$$\hat{\mu}^{*} = \frac{1}{7}\left[\frac{y_{11}}{1} + \frac{y_{21}}{1/2} + \frac{(y_{32}+y_{33})}{1} + \frac{(y_{42}+y_{43})/2}{1/2}\right]$$
$$= \frac{1}{7}\left[\frac{2}{1} + \frac{1}{1/2} + \frac{(62+66)}{1} + \frac{(47+3)/2}{1/2}\right] = 26.$$

Simulation 2: Balanced Probability-based LHD

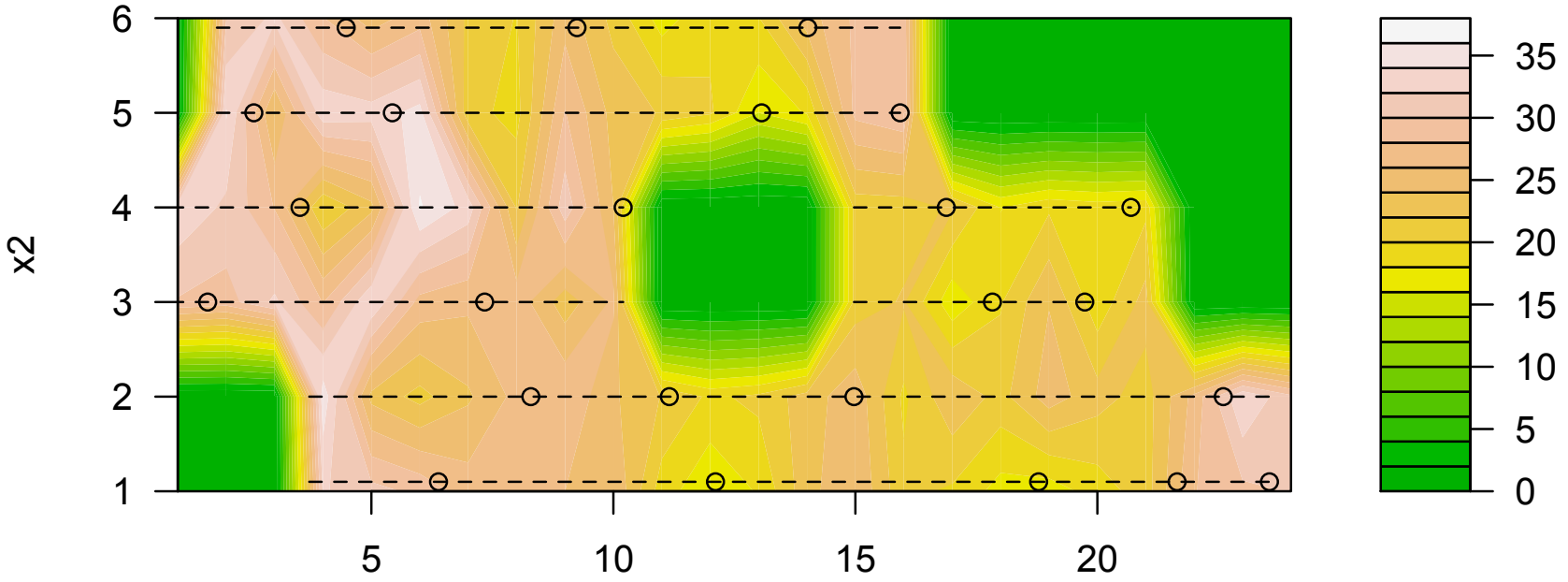
		0 (2/7)	10 (2/7)	3 (3/7)	0 (7/7)
	2 (4/7)	8 (3/7)	24 (3/7)	25 (4/7)	
1 (7/7)	0 (3/7)	0 (2/7)	4 (2/7)		

- All possible adaptive BPLHDs and a comparison of the unbiased estimators

sampler	$\hat{\mu}_{\mathrm{B}}$	$\hat{\mu}_{\mathrm{B}}^{\mathrm{RB}}$	$\hat{\mathrm{var}}(\hat{\mu}_{\mathrm{B}})$	$\hat{\mathrm{var}}(\hat{\mu}_{\mathrm{B}}^{\mathrm{RB}})$
$\{11\},\{21\},\{32\},\{42\},\{53\},\{63\};\{52\},\{43\},\{41\}$	6.99	8.15	2.23	5.36
$\{11\},\{21\},\{32\},\{43\},\{52\},\{63\};\{42\},\{41\},\{53\}$	9.32	8.15	11.22	5.36
$\{11\},\{21\},\{33\},\{42\},\{52\},\{63\};\{32\},\{41\},\{43\},\{53\}$	4.85	4.85	3.24	3.24
$\{11\},\{22\},\{31\},\{42\},\{53\},\{63\};\{52\},\{32\},\{41\},\{43\}$	5.72	6.89	2.50	5.40
$\{11\},\{22\},\{31\},\{43\},\{52\},\{63\};\{42\},\{32\},\{41\},\{53\}$	8.06	6.89	11.02	5.40
$\{11\},\{22\},\{32\},\{41\},\{53\},\{63\}$	3.68	3.68	3.11	3.11
$\{11\},\{22\},\{33\},\{41\},\{52\},\{63\};\{42\},\{32\},\{43\},\{53\}$	6.42	6.31	0.23	0.23
mean	6.42	6.42	4.79	4.02

Data Center Example

- 24 initial sensors in the initial design



- Comparison of adaptive design with non-adaptive design (simple random sampling with same sample size)

	$\hat{\mu}$	$\hat{\mu}^*$	$\hat{\mu}_{\text{SRS}}$
mean	24.20	24.20	24.21
variance	0.53	0.47	0.78

Conclusion

- Most of the existing space-filling designs are constructed for rectangular regions.
- A new class of space-filling designs is introduced for slid-rectangular regions.
- Adaptive designs and unbiased estimators are discussed.
- Comparisons between adaptive designs and non-adaptive designs are performed. It appears that adaptive designs can reduced estimation variations.
- Ongoing work: theoretical derivations, ratio and regression estimators, etc.

Thank you!!