Stat 890
Design of computer experiments

- Will introduce design concepts for computer experiments
- Will look at more elaborate constructions next day

SFU
Department of Statistics and Actuarial Science
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Experiment design

- In computer experiments, as in many other types of experiments considered by statisticians, there are usually some **factors** that can be adjusted and some **response variable** that is impacted by (some) of the factors

- The experiment is conducted, for example to see which/how the factors impact the response

- Generally, the **experimental design** is the set of factor level combinations, or treatments, that is applied to experimental units (treatment structure) and as well as the mechanism for assigning treatments to experimental units (randomization structure)

- For deterministic computer models, the randomization structure is not important… neither is replication

Experiment design

- For a computer experiment, the experiment design is the set of d-dimensional inputs where the computational model is evaluated

- Notation : X is an n x d, design matrix; $y(X)$ is the n x 1 vector of responses

- Experimental region is usually the unit hypercube $[0,1]^d$

Experiment design

- **Experiment goals:**
  - Computing the mean response
  - Response surface estimation or computer model emulation
  - Sensitivity analysis
  - Optimization
  - Estimation of contours and percentiles
  - .
  - .
  - .

Design for estimating the mean response

- Suppose have a computer model *y(*x), where x is a *d*-dimensional uniform random variable on $[0,1]^d$

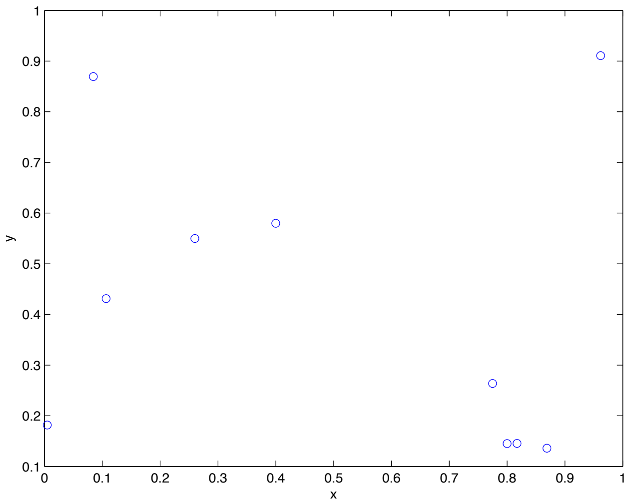
- Interested in $\mu = \int y(x)dx$

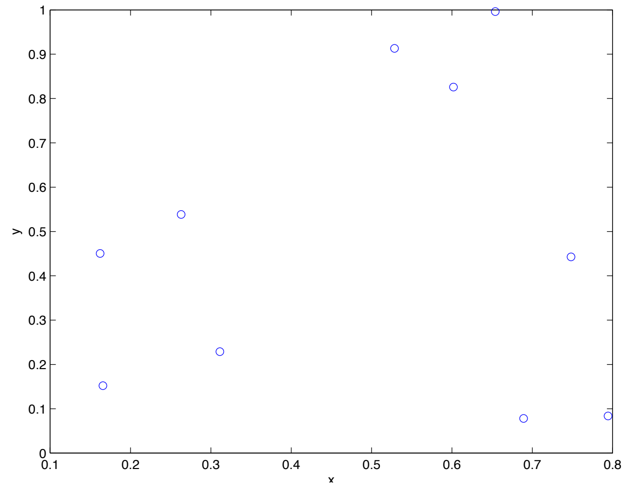
- One way to do this is to randomly sample x of size *n* from $U[0,1)^d$ (Monte Carlo method)

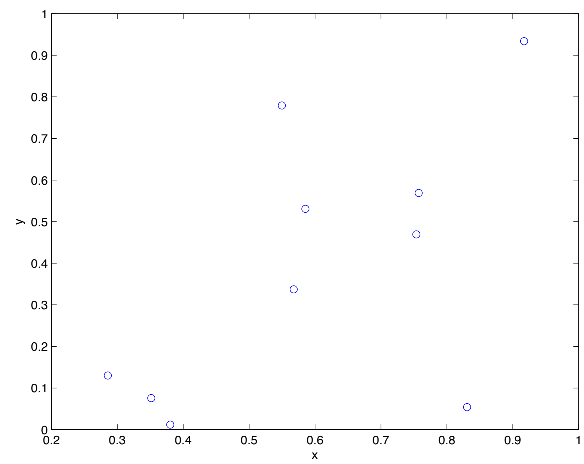
$$\hat{\mu} = \frac{\sum_{i=1}^{n} y(x_i)}{n}$$

SFU
Department of Statistics and Actuarial Science
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Design for estimating the mean response







Experiment design

- **Good news:** approach gives an unbiased estimator of the mean
- **Bad news:** approach can often lead of designs that miss much of the input space… relatively large variance for the estimator
- Could imagine taking a regular grid of points as the design
  - Impractical since need many points in moderate dimensions

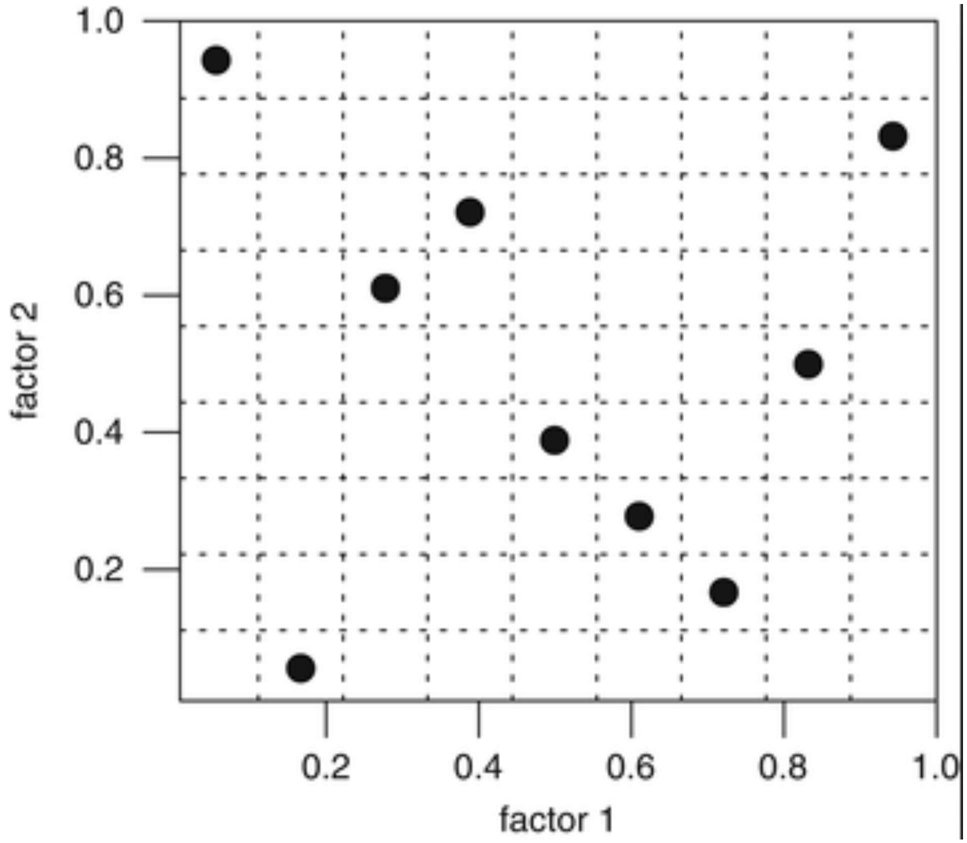
Experiment design

- McKay, Beckman and Conover (1979, *Technometrics*) introduced **Latin hypercube sampling** as an approach for estimating the mean

- Random Latin hypercube design in 2-d (sample size = n)
  - Construct an n x n grid over the unit square
  - Construct an n x 2 matrix, Z, with columns that are independent permutations of the integers {1, 2, …, n}
  - Each row of the matrix is the index for a cell on the grid. For the i^{th} row of Z, take a random uniform draw from the corresponding cell in the n x n grid over the unit square… call this x_i (i=1,2,…n)

SFU
Department of Statistics and Actuarial Science
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Design for estimating the mean response



Enforces a 1-*d* projection property

Is an attempt at filling the space

Easy to construct

Can get some pretty bad designs (what happens if the permutations result in column 1=column 2?)

Experiment design

- McKay, Beckman and Conover (1979, *Technometrics*) introduced **Latin hypercube sampling** as an approach for estimating the mean

$$\mathrm{Var}(\overline{Y}_R(t)) = (1/N)\,\mathrm{Var}(Y(t))$$

$$\mathrm{Var}(\overline{Y}_S(t)) = \mathrm{Var}(\overline{Y}_R(t)) - (1/N^2) \sum_{i=1}^{N} (\mu_i - \mu)^2$$

$$\mathrm{Var}(\overline{Y}_L(t)) = \mathrm{Var}(\overline{Y}_R(t)) + ((N-1)/N) \cdot 1/(N^K(N-1)^K)) \sum_R (\mu_i - \mu)(\mu_j - \mu) \quad (3.2)$$

where $\mu = E(Y(t))$,

$\mu_i = E(Y(t) \mid \mathbf{X} \in S_i)$ in the stratified sample, or

$\mu_i = E(Y(t) \mid \mathbf{X} \in \text{cell } i)$ in the Latin hypercube sample,

Experiment design

- **Good news:** lower variance than random sampling
- **Bad news:** can still get large holes in the input space
- What properties would you like to see to improve the design?

Experiment design

- **Space-filling criteria:**
  - **Maximin designs:** For a design X, maximize the minimum distance between any two points
  - **Minimax designs**: minimizes the max distance between any point in the input region and the points in the design X
  - Can find designs that optimize one of these criteria or apply the criteria to a class of designs (e.g., Latin hypercube designs)
  - (we will construct some of these on the next assignment)

Experiment design

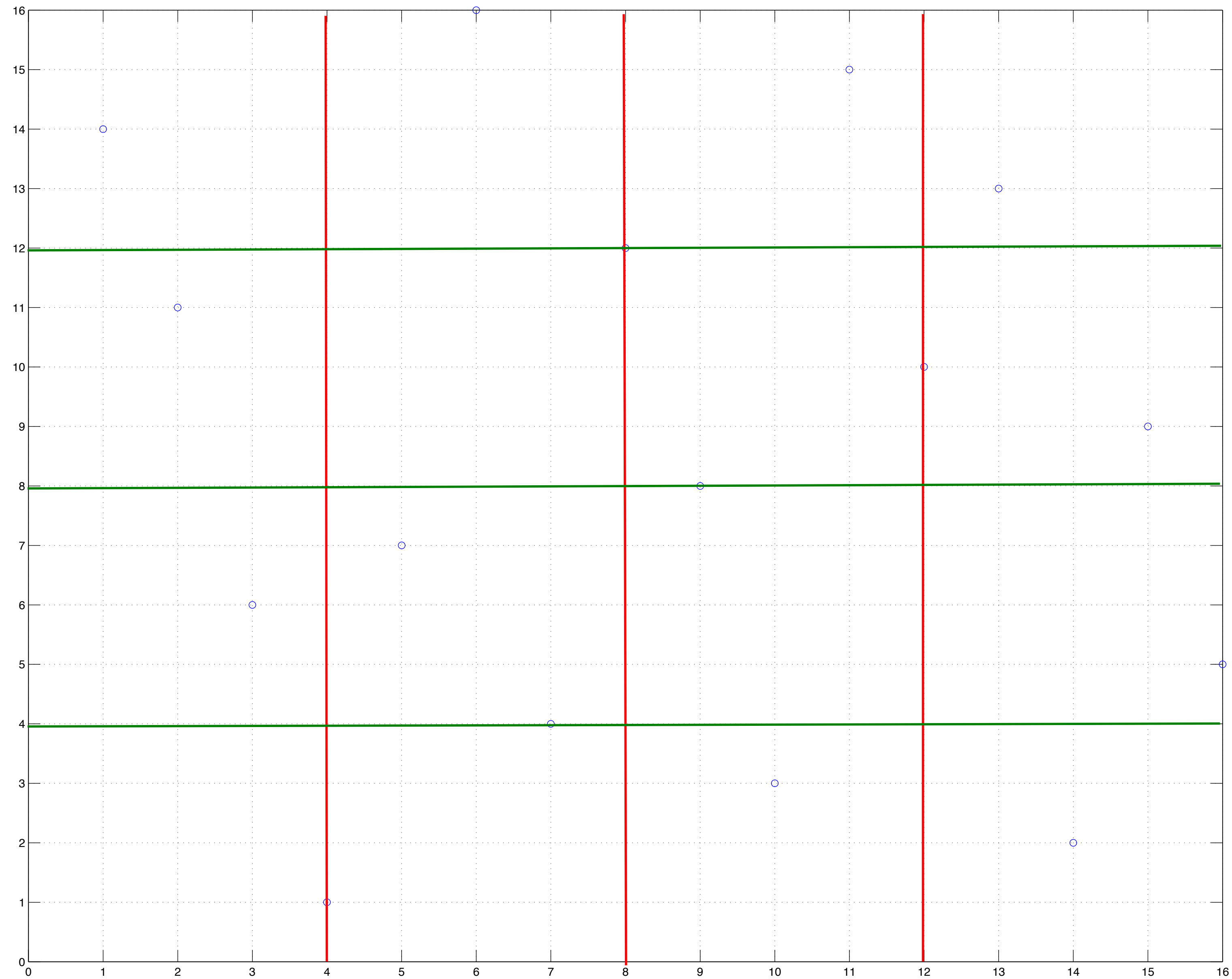
- **Space-filling criteria:**
  - Orthogonal array based Latin hypercubes (Tang, 1993)
  - This is adds restriction on the random permutation of the integers {1,2, …, n} when constructing the Latin hypercube
  - **Idea:** start with an orthogonal array (array with s>2 symbols per column and all possible combinations of the s symbols in the rows of the design appear equally often)
  - Restrict a permutation of consecutive integers to those rows that have a particular symbol (each row gets a unique value in {1,2, …, n})

Experiment design

1	1
1	2
1	3
1	4
2	1
2	2
2	3
2	4
3	1
3	2
3	3
3	4
4	1
4	2
4	3
4	4

→

4	1
3	6
2	11
1	14
7	4
5	7
8	12
6	16
10	3
9	8
12	10
11	15
14	2
16	5
15	9
13	13


0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Experiment design

- **Good news:** Tang showed that such designs can achieve smaller variance than LHS for estimating the mean

- **Bad news:** Orthogonal arrays do not exist for all run sizes (for 2 symbol designs, the run sizes are powers of 4)

- The design we just looked at is a 4^2 grid

- Strength 2 orthogonal array has all pairs of columns that have the combinatorial property, but do not need all triplets to have the same property (think fractional factorial designs)

- Is particularly useful if the design is dominated by two of the inputs (*factor sparsity*)

Experiment design

- **Good news:** Tang showed that such designs can achieve smaller variance than LHS for estimating the mean
- **Bad news:** Orthogonal arrays do not exist for all run sizes (for 2 symbol designs, the run sizes are powers of 4)
- The design we just looked at is a 4^2 grid
- Strength 2 orthogonal array has all pairs of columns that have the combinatorial property, but do not need all triplets to have the same property (think fractional factorial designs)
- Is particularly useful if the design is dominated by two of the inputs (*factor sparsity*)

Comment

- The justification for the designs discussed so far are based on the practical problem of estimating the mean
- There is an entire area of mathematics that also considers this problem - *quasi-Monte Carlo*
- However, the designs considered so far are often used to *emulate* the computer model (i.e., the design that is run, followed by fitting a GP)
- Why do you suppose that is?

Designs based on distance

- Clearly space filling is an important property generally
- What about for computer model emulation?
- Consider an n point design over [0,1]
- To avoid points being too close, can use a criterion that aims to spread out the points

Designs based on distance

- **Maximin designs:** the idea for a maximin is to maximize the minimum distance between any two design points

- The distance criterion is usually written:

$$c_p(x_1, x_2) = \left[\sum +_{j=1}^{d} |x_{1j} - x_{2j}|^p\right]^{1/p}$$

- The *best* design is the collection of *n* distinct points from $[0,1]^d$ where this is maximized over the design

- **How would you get one in practice?**

Designs based on distance

- **Minimax designs:** when you are making a prediction using, say, a GP, would like to have nearby design points. So, at the potential prediction points, would like to minimize the maximum distance

- The *best* design is the collection of *n* distinct points from $[0,1]^d$ where the maximum distance from any point in the design region is small as possible

$$\min_{X \in [0,1]} \max_{x \in [0,1]} c_p(x, X)$$

- You might have to do this on an assignment!

Designs based on distance

- Johnson, Moore and Ylvisaker (1990, *JSPI*) proposed the use of these designs for computer experiments
- Developed asymptotic theory under which both of the sorts of design can be optimal… also relate the theory to D-optimal designs
- An important paper worth understanding
- Kunsch et al. (2005, *IEEE T. on Inf. Theory*) develops complementary theory for lattices

Designs based on distance

- **Good news:** criteria give designs that have intuitively appealing space filling properties
- **Bad news:** In practice, the designs are not always good
- Maximin designs frequently place points near the boundary of the design space
- Minimax designs are often very hard to find

Combining criteria

- As before, good idea to combine criteria – e.g., maximin Latin hypercube
- Other approaches include considering Latin hypercube designs where the columns of the design matrix have low correlation (e.g., Owen 1994; Tang 1998)