Experimental Design part 1

Advanced Systems Lab – 263-0007-00

G. Alonso, D. Kossmann

Systems Group, ETH Zürich, Switzerland

2011

Introduction

What does it mean to design an experiment?

Introduction

What does it mean to design an experiment?
Performance is affected by a large number of factors

- Workloads
- Systems
- Knobs

Introduction

What does it mean to design an experiment?
Performance is affected by a large number of factors

- Workloads
- Systems
- Knobs

We are interested in:

- Which ones are the most important?
- Which ones are related?

Goal: get the most information with least effort (minimum number of experiments)

Response variable

A response variable is the outcome of an experiment - typically the measured performance of the system. E.g.:

- task throughput
- response time
- . . . or any other metric (not just performance)

Factors

A factor is any variable that affects the response, and which has several alternatives.

- A factor is a "knob" in the experiment.
- Sometimes also called "predictor variables" or "predictors".
- E.g. queue length, buffer size, CPU type, etc.

Levels

Levels are the values that a given factor can assume - the alternatives for a factor. E.g.

- Buffer size in kilobytes
- Different choices of disk interface
- Different database schemas
- Clock speeds of a processor
- ...

Primary and Secondary Factors

- Primary factors are those whose effects need to be quantified
  - E.g. what is the effect of replication factor on performance?

Primary and Secondary Factors

- Primary factors are those whose effects need to be quantified
  - E.g. what is the effect of replication factor on performance?
- Secondary factors are those that impact performance but whose effect we are not interested in quantifying
  - E.g. server hardware performance for your PostgreSQL trials

Replication

Replications are the number of times each experiment is to be repeated with particular levels for each factor.

Design

An experimental design consists of:

Design

An experimental design consists of:

- the number of different experiments
- the factor level combinations for each experiment
- the number of replications of each experiment

Experimental Unit

An experimental unit is any entity used for the experiment

- E.g. server + PostgreSQL + network

We are not interested in comparing the experimental units against each other ⇒ a goal of the design is to minimize the impact of variation between the units

Interaction

Two factors interact if the effect of one depends on the level of the other.

Interaction considerably complicates the business of interpreting experimental results

Interacting and non-interacting factors

Non-interacting:

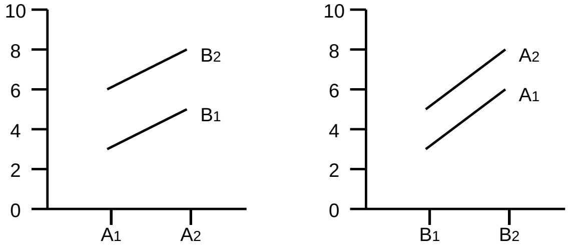
	A_1	A_2
B_1	3	5
B_2	6	8

Interacting:

	A_1	A_2
B_1	3	5
B_2	6	9

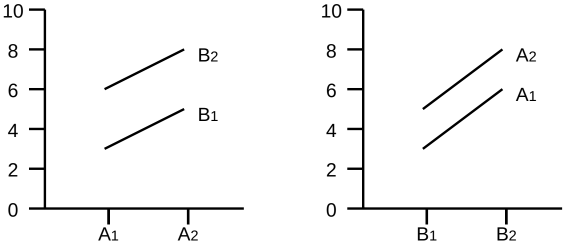
Interacting and non-interacting factors

Non-interacting:

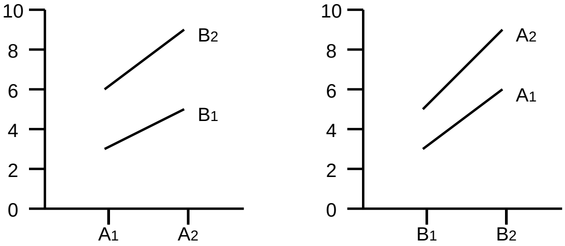


Interacting and non-interacting factors

Non-interacting:



Interacting:



Common mistakes

Try to avoid the following:

- Ignoring the variation due to experimental errors
- Not controlling important parameters (secondary factors)
- Not isolating the effects of different factors
- Overly simple (and very inefficient designs)
- Ignoring interactions between factors
- Conducting too many experiments
  - Take it slowly!
  - Break up the project into steps

(Too) simple designs

- Pick a typical configuration
- Vary one factor at a time
- Find out which value is best, and fix it
- Repeat for each factor

How much work?
For k factors, n_i levels per factor, #experiments needed is:

$$n = 1 + \sum_{i=1}^{k} (n_i - 1)$$

(Too) simple designs

Don't do this!

1. It's not statistically efficient.
   We do much better in reducing effort (see later)
2. It doesn't always work!
   If factors interact, search can find a local maximum.

	A			
	1	2	3	4
B	1	5	5	6
	1	2	3	4
	1	1	3	8

(Too) simple designs

Don't do this!

1. It's not statistically efficient.
   We do much better in reducing effort (see later)
2. It doesn't always work!
   If factors interact, search can find a local maximum.

	A			
	1	2	3	4
B	1	*5*	5	**6**
	1	2	3	4
	1	1	3	8

Full factorial designs

- Try all combinations of factors and levels.
- For k factors, n_i levels per factor, #experiments needed is:

$$n = \prod_{i=1}^{k} n_i$$

Full factorial designs

- Try all combinations of factors and levels.
- For k factors, n_i levels per factor, #experiments needed is:

$$n = \prod_{i=1}^{k} n_i$$

- This is a **lot** of work!
  (consider the configuration options for PostgreSQL. . .)

Full factorial designs

- Try all combinations of factors and levels.
- For k factors, n_i levels per factor, #experiments needed is:

$$n = \prod_{i=1}^{k} n_i$$

- This is a **lot** of work!
  (consider the configuration options for PostgreSQL...)
- Options to reduce the search space:
  1. Reduce the number of levels for each factor
  2. Reduce the number of factors
  3. Use fractional factorial designs

2^k Factorial Designs

Reduce the number of levels for each of k factors to 2

Particularly handy for:

- My system is better than your system
- Feature X improves system Y

Requires each factor to be unidirectional or monotonic

Example: 2^2 Factorial Design

Observation: can update book examples by multiplying by 1000!

	Memory size	
Cache size (MB)	4GB	16GB
1	15	45
2	25	75

Define variables x_A and x_b to represent levels for each factor:

$$x_A = \begin{cases} -1 & \text{if 4 GB main memory;} \\ 1 & \text{if 16 GB main memory.} \end{cases}$$

$$x_B = \begin{cases} -1 & \text{if 1 MB cache,} \\ 1 & \text{if 2 MB cache.} \end{cases}$$

Example: 2^2 Factorial Design

A useful fiction: non-linear regression model for performance:

$$y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B$$

Example: 2^2 Factorial Design

A useful fiction: non-linear regression model for performance:

$$y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B$$

This means we can write:

$$\begin{aligned}
15 &= q_0 - q_A - q_B + q_{AB} \\
45 &= q_0 + q_A - q_B - q_{AB} \\
25 &= q_0 - q_A + q_B - q_{AB} \\
75 &= q_0 + q_A + q_B + q_{AB}
\end{aligned}$$

Example: 2^2 Factorial Design

Solving:

$$y = 40 + 20x_A + 10x_B + 5x_Ax_B$$

What does this mean?

Example: 2^2 Factorial Design

Solving:

$$y = 40 + 20x_A + 10x_B + 5x_Ax_B$$

What does this mean?

- Mean performance is 40
- Effect of memory is 20
- Effect of cache is 10
- Interaction between the two accounts for 5

Sign table calculation

The same calculation can be done using a sign table:

I	A	B	AB	y
1	-1	-1	1	15
1	1	-1	-1	45
1	-1	1	-1	25
1	1	1	1	75
160	80	40	20	Total
40	20	10	5	Total/4

Allocation of variation

How to measure the importance of a factor

We can measure the importance of a factor by the proportion of the total variation that it is reponsible for.

Sample variance is:

$$y = s_y^2 = \frac{\sum_{i=1}^{2^2} (y_i - \bar{y})^2}{2^2 - 1}$$

where $\bar{y}$ is the mean of responses from all 4 experiments.

Allocation of variation

How to measure the importance of a factor

Total variation of y is the Sum of Squares Total (SST):

$$SST = \sum_{i=1}^{2^2} (y_i - \bar{y})^2$$

Note: variation $\neq$ variance!

For a 2^2 design, this is:

$$SST = 2^2 q_A^2 + 2^2 q_B^2 + 2^2 q_{AB}^2$$

or

$$SST = SSA + SSB + SSAB$$

So fraction of variation explained by $A = \frac{SSA}{SST}$

Allocation of variation

How to measure the importance of a factor

For proof of this, see the book.

For our example:

$$\bar{y} = \frac{1}{4}(15 + 55 + 25 + 75) = 40$$

$$Total\ variation = \sum_{i=1}^{4} (y_i - \bar{y})^2 = (25^2 + 15^2 + 15^2 + 35^2)$$

$$= 2100 = 4 \times 20^2 + 4 \times 10^2 + 4 \times 5^2$$

Allocation of variation

How to measure the importance of a factor

For our example:

Total variation:	2100	100%
Memory:	1600	76%
Cache:	400	19%
Both:	100	5%

$2^k r$ Factorial design with replication

- 2^k factorial designs don't repeat any experiment
  $\Rightarrow$ no estimate of experimental error

$2^k r$ Factorial design with replication

- 2^k factorial designs don't repeat any experiment
  $\Rightarrow$ no estimate of experimental error
- Solution is to repeat every trial r times
- We add an error term to the model:

$$y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B + e$$

$2^k r$ Factorial design with replication

Consider the previous experiment, but with 3 trials each:

I	A	B	AB	y	$\bar{y}$
1	-1	-1	1	(15, 18, 12)	15
1	1	-1	-1	(45, 48, 51)	48
1	-1	1	-1	(25, 28, 19)	24
1	1	1	1	(75, 75, 81)	77
164	86	38	20		Total
41	21.5	9.5	5		Total/4

$2^k r$ Factorial design with replication

We now have the q_i values, but we cannot allocate variation as before, because we don't actually know the "true" response (it's a random variable!).

The estimated response $\hat{y}_i$ for experiment i, where $A = x_{A_i}$ and $B = x_{B_i}$, is:

$$\hat{y}_i = q_0 + q_A x_{A_1} + q_B x_{B_i} + q_{AB} x_{A_i} x_{B_i}$$

$2^k r$ Factorial design with replication

The experimental error in trial j of experiment i is the difference between this estimate and the actual result:

$$e_{ij} = y_{ij} - \hat{y}_i = y_{ij} - q_0 - q_A x_{A_1} - q_B x_{B_i} - q_{AB} x_{A_i} x_{B_i}$$

$2^k r$ Factorial design with replication

The experimental error in trial j of experiment i is the difference between this estimate and the actual result:

$$e_{ij} = y_{ij} - \hat{y}_i = y_{ij} - q_0 - q_A x_{A_1} - q_B x_{B_i} - q_{AB} x_{A_i} x_{B_i}$$

The sum of all the errors must be zero. The sum of the squared errors (SSE) can be used to estimate the variance of the errors and compute confidence intervals for the effects:

$$SSE = \sum_{i=1}^{2^2} \sum_{j=1}^{r} e_{ij}^2$$

$2^k r$ Factorial design with replication

Estimating errors in the sign table:

	Effect				Estimated	Measured					
	I	A	B	AB	Response	Responses			Errors		
i	41	21.5	9.5	5	$\hat{y}_i$	y_{i1}	y_{i2}	y_{i2}	e_{i1}	e_{i2}	e_{i3}
1	1	-1	-1	1	15	15	18	12	0	3	-3
2	1	1	-1	-1	48	45	48	51	-3	0	3
3	1	-1	1	-1	24	25	28	19	1	4	-5
4	1	1	1	1	77	75	75	81	-2	-2	4

We get SSE = 102 (try it!)

Allocation of variation for $2^k r$

Total sum of squares is given by:

$$SST = \sum_{i,j} (y_{ij} - \bar{y}_{..})^2$$

Which has the following parts (see book for derivation):

$$\sum_{i,j} (y_{ij} - \bar{y}_{..})^2 = 2^2 r q_A^2 + 2^2 r q_B^2 + 2^2 r q_{AB}^2 + \sum_{i,j} e_{ij}^2$$

$$SST = SSA + SSB + SSAB + SSE$$

- where SSE is the variation due to experimental error.

Allocation of variation for $2^k r$

For our example:

$$
\begin{aligned}
SSY &= 15^2 + 18^2 + 12^2 + 45^2 + \ldots + 75^2 + 75^2 + 81^2 = 27204 \\
SS0 &= 2^2 r q_0^2 = 12 \times 41^2 = 20172 \\
SSA &= 2^2 r q_A^2 = 12 \times (21.5)^2 = 5547 \\
SSB &= 2^2 r q_B^2 = 12 \times (9.5)^2 = 1083 \\
SSAB &= 2^2 r q_{AB}^2 = 12 \times 5^2 = 300 \\
SSE &= 27204 - 2^2 \times 3(41^2 + 21.5^2 + 9.5^2 + 5^2) = 102 \\
SST &= SSY - SS0 = 27204 - 20172 = 7032
\end{aligned}
$$

Note that (as expected):

$$SSA + SSB + SSAB + SSE = SST$$

Allocation of variation

For our example:

Total variation:	7032	100%
Memory:	5547	78.88%
Cache:	1083	15.4%
Both:	300	4.27%
Errors:	102	1.45%

Assumptions

What are we assuming in all this analysis?

- Model errors are IID normal, with zero mean and constant standard deviation
  - or very small, relative to effects
- Errors are additive
- Effects of factors are additive

Make sure these assumptions hold! See the book for various tests you can use to confirm this.