

# **2001: Dissertation Process**

## **Measurement in data-poor situations**

**Dr. Mathias (Mat) Disney**

**UCL Geography**

**Office: 113 Pearson Building**

**Tel: 7670 0592**

**Email: [mdisney@ucl.geog.ac.uk](mailto:mdisney@ucl.geog.ac.uk)**

**[www.geog.ucl.ac.uk/~mdisney](http://www.geog.ucl.ac.uk/~mdisney)**

# Overview

- What do we mean by data-poor?
- Types of measurement: asking the right question
- Types of sampling: looking in the right place
- Statistical testing, modelling and parsimony: making best use of what you have

**WRITE DOWN A NUMBER BETWEEN 1:365**

# What do we mean by data-poor?

- Few measurements or observations
  - Fewer than perhaps we would like?
- Few data not necessarily a problem
- e.g. if I want to know how tall I am
  - How many measurements do I need?
    - How accurate are my measurements?
    - How accurate do I want/need to be?
    - How do I express uncertainty in my measurements & answer?

# What do we mean by data-poor?

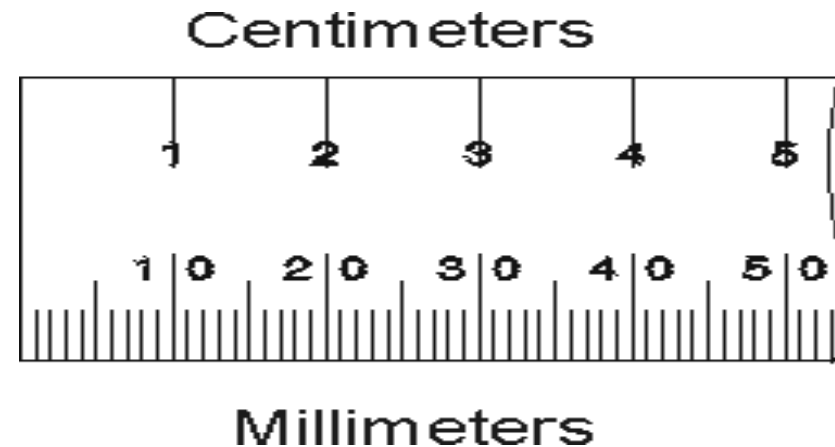
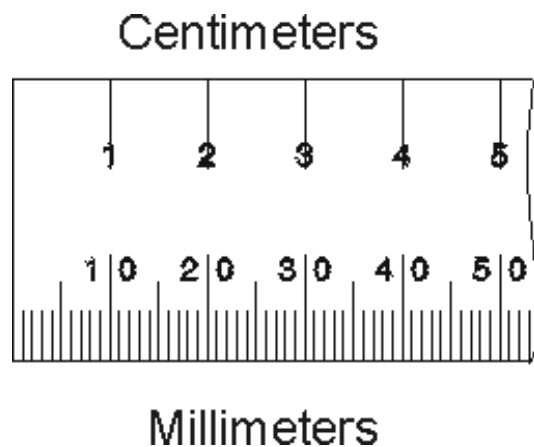
- Examples
  - Average height of a group (sample) of people from a larger group (population) - how many do I measure? 10? 20?
  - E.g. how many people in this room?
  - Is this sample “representative”?
- What if I ask a more difficult question?
  - E.g. Do you approve of Government’s policy on tuition fees?
  - Is a yes/no/don’t know answer helpful?
  - How do I quantify any sources of error now?
  - Who do I put the question to?

# We are data-poor when....

- We have small number of samples (see random errors) and/or selection bias (see systematic errors) and/or limited time/resources e.g.
  - Questionnaires on hard-to-measure socio-economic indicators
  - Measurements of highly variable systems
- We have large samples BUT large variation e.g.
  - Temperature data over UK
  - Incidences of a particular type of cancer
- It is hard/impossible to measure variables we are interested in directly
  - e.g. Climate change? Voting intention?

# Errors and uncertainty

- Random errors
  - Examples
    - Physical measurement of distance, time, mass, velocity, voltage
    - Any instrument/operator has a **precision**
    - NOT the same as accuracy!



# Errors and uncertainty

- Random errors are easy (ish) to deal with
  - Take several/many measurements (sample “true” value) to give a mean value PLUS some estimate of uncertainty
  - $\alpha$  is standard error of mean;  $\sigma$  is standard deviation;  $N$  is number of samples
  - Quote our result as mean  $\pm \alpha$
  - So, typically reduce  $\alpha$  by  $1/\sqrt{N}$

$$\alpha = \frac{\sigma_{N-1}}{\sqrt{N}}$$

- Useful link
  - <http://level1.physics.dur.ac.uk/skills/erroranalysis.php>

# Errors and uncertainty

- **Systematic errors**
  - Offset or bias in measurements (can be constant or variable)
  - Harder to deal with and must be identified with care
  - E.g. Wrongly-calibrated instrument
  - Making measurements consistently but incorrectly
- **Particularly problematic for survey data**
  - Is a sample “representative”? What do we mean by “representative”? Is there selection bias?
  - **MUST** think v. carefully about possible bias & EXPLICITLY consider/remove selection bias in experimental design
  - <http://instructor.physics.lsa.umich.edu/int-labs/Statistics.pdf>



# Errors and uncertainty

- The ‘Gold Standard’: randomised double blind
  - Single group divided into two samples by e.g. tossing coin (random assignment to group A or B)
  - Sample A treated in some way
  - Sample B given placebo
  - Neither researchers nor participants know which is which until study ends (both “blind”).

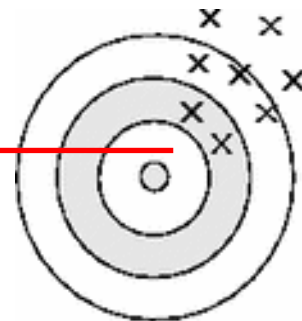
Goldacre, B. (2009) *Bad Science*, Harper Perennial, pp 288.

[www.badscience.net](http://www.badscience.net)

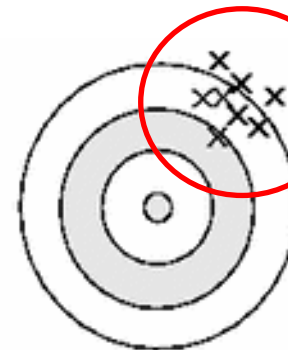
[http://en.wikipedia.org/wiki/Randomized\\_controlled\\_trial](http://en.wikipedia.org/wiki/Randomized_controlled_trial)

# Errors and uncertainty: summary

Offset is (probably!)  
systematic error

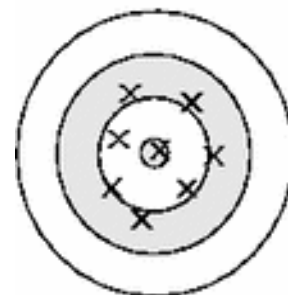


Not precise  
Not accurate

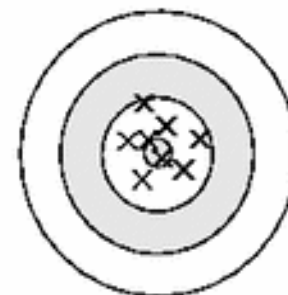


Precise  
Not accurate

Spread is (probably!)  
random error



Not precise  
Accurate



Precise  
Accurate

Figure from: <http://www.mathworks.com/access/helpdesk/help/toolbox/daq/?/access/helpdesk/help/toolbox/daq/f5-28876.html> & <http://www.google.co.uk/search?hl=en&q=precision+v+accuracy&meta=>

## Asking the right question

- What response are you expecting and why?
- Is the measurement you make the “best” one, given your hypothesis?
- If not, why? Can you find a better one?
- Have you phrased your experiment/hypothesis in such a way as to make it testable logically?

# Example from polling

- See <http://www.populuslimited.com/>
  - “within each government office region a random sample of telephone numbers was drawn from the entire BT database of domestic telephone numbers. Each number so selected had its last digit randomised so as to provide a sample including both listed and unlisted numbers”
  - Issues/Assumptions?
  - See Benford’s Law and Zipf’s Law
    - <http://mathworld.wolfram.com/BenfordsLaw.html>
    - [http://www.cut-the-knot.org/do\\_you\\_know/zipfLaw.shtml](http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml)

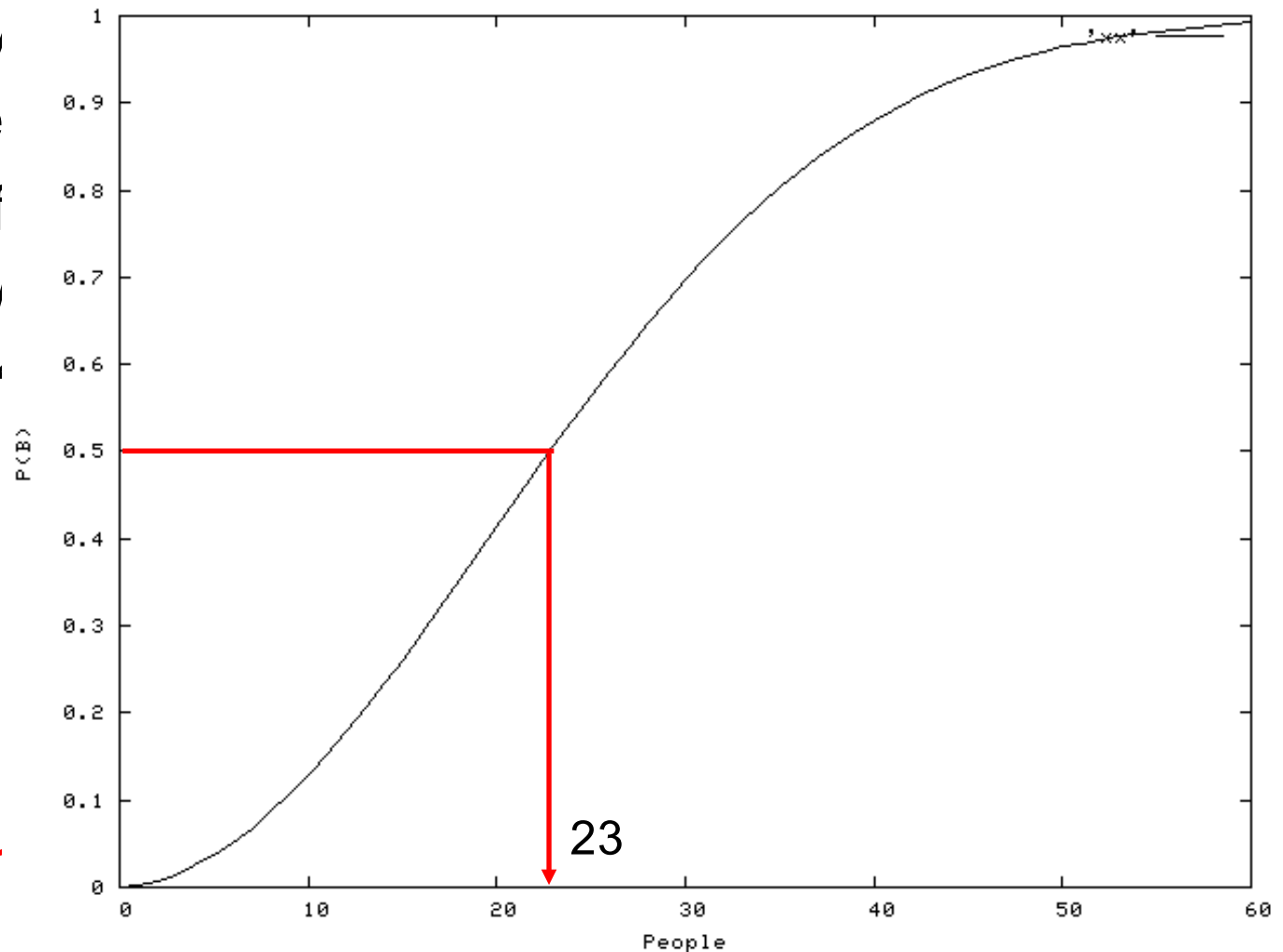
# Example from polling

- From <http://www.populuslimited.com/>
- Problems/bias??
  - “Data were weighted to the profile of all adults aged 18+ (including non telephone owning households). Data were weighted by sex, age, social class, household tenure, work status, number of cars in the household and whether or not respondent has taken a foreign holiday in the last 3 years. Targets for the weighted data were derived from the National Readership survey, a random probability survey comprising 34,000 random face-to-face interviews conducted annually.”
  - Issues/Assumptions?

# Probability is a funny thing

- How many people do I need in a room before  $P(B_{a,b})$  the probability birthday, is be
- i.e. what is  $N$  if
- $P(B_{a,b}) = 365!$
- The Birthday “
- Need to be careful about relying on intuition!
- NB assumes all birthdays equally likely.....

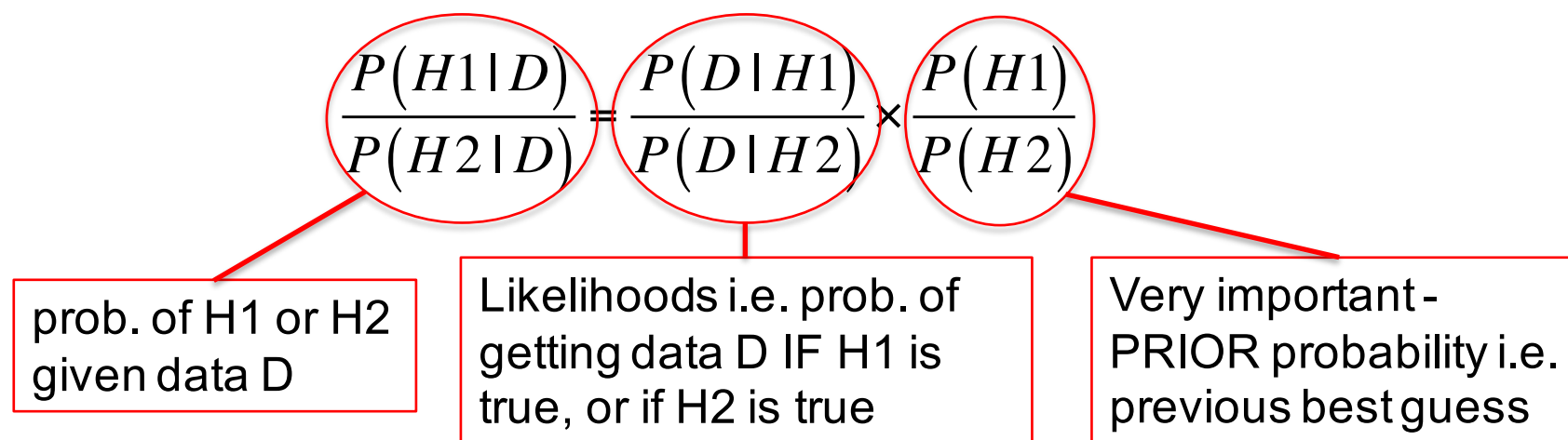
**Look at the nur**



# The tragic case of Sally Clark

- Two cot-deaths (SIDS), 1 year apart, aged 11 weeks and 8 weeks. Mother Sally Clark charged with double murder, tried and convicted in 1999
  - Statistical evidence was misunderstood, “expert” testimony was wrong, and a fundamental logical fallacy was introduced
- What happened?
- We can use Bayes’ Theorem to decide between 2 hypotheses
  - H1 = Sally Clark committed double murder
  - H2 = Two children DID die of SIDS
- <http://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/>
- <http://yudkowsky.net/rational/bayes>

# The tragic case of Sally Clark



- Data? We observe there are 2 dead children
- We need to decide which of H1 or H2 are more plausible, given D (and prior expectations)
- i.e. want ratio  $P(H1|D) / P(H2|D)$  i.e. odds of H1 being true compared to H2, GIVEN data and prior



# The tragic case of Sally Clark

- **ERROR 1: events NOT independent**
- $P(1 \text{ child dying of SIDS})? \sim 1:1300$ , but for affluent non-smoking, mother  $> 26$  yrs  $\sim 1:8500$ .
- Prof. Sir Roy Meadows (expert witness)
  - $P(2 \text{ deaths})? 1:8500 \times 8500 \sim 1:73 \text{ million}$ .
  - This was KEY to her conviction & is demonstrably wrong
  - $\sim 650000$  births a year in UK, so at 1:73M a double cot death is a 1 in 100 year event. BUT 1 or 2 occur every year – how come?? No one checked ...
  - NOT independent  $P(2^{\text{nd}} \text{ death} | 1^{\text{st}} \text{ death})$  5-10 higher i.e. 1:100 to 200, so  $P(H_2)$  actually  $1:1300 \times 5/1300 \sim \underline{1:300000}$

# The tragic case of Sally Clark

- **ERROR 2: “Prosecutor’s Fallacy”**
  - 1:300000 still VERY rare, so she’s unlikely to be innocent, right??
    - Meadows “Law”: ‘one cot death is a tragedy, two cot deaths is suspicious and, until the contrary is proved, three cot deaths is murder’
  - **WRONG**: Fallacy to mistake chance of a rare event as chance that defendant is innocent
    - In large samples, even rare events occur quite frequently - someone wins the lottery (1:14M) nearly every week
    - 650000 births a year, **expect** 2-3 double cot deaths.....
    - AND we are ignoring rarity of double murder (H1)

# The tragic case of Sally Clark

- **ERROR 3: ignoring odds of alternative (also very rare)**
  - Single child murder v. rare (~30 cases a year) BUT generally significant family/social problems i.e. NOT like the Clarks.
  - $P(1 \text{ murder}) \sim 30:650000$  i.e. 1:21700
  - Double MUCH rarer, BUT  $P(2^{\text{nd}}|1^{\text{st}} \text{ murder}) \sim 200 \text{ x more likely given first, so } P(H1|D) \sim (1/21700 * 200/21700) \sim 1:2.4\text{M}$
- So, two very rare events, but double murder  $\sim 10 \text{ x rarer than double SIDS}$
- **So  $P(H1|D) / P(H2|D)$ ?**
  - $P(\text{murder}) : P(\text{cot death}) \sim 1:10$  i.e. 10 x more likely to be double SIDS
  - Says nothing about guilt & innocence, just relative probability

# The tragic case of Sally Clark

- Sally Clark acquitted in 2003 after 2<sup>nd</sup> appeal (but not on statistical fallacies) after 3 yrs in prison, died of alcohol poisoning in 2007
  - Meadows “Law” redux: triple murder v triple SIDS?
    - In fact,  $P(\text{triple murder} \mid 2 \text{ previous}) : P(\text{triple SIDS} \mid 2 \text{ previous}) \sim ((21700 \times 123) \times 10) / ((1300 \times 228) \times 50) = 1.8:1$
    - So  $P(\text{triple murder}) > P(\text{SIDS})$  but not by much
- Meadows’ ‘Law’ should be:
  - ‘when three sudden deaths have occurred in the same family, statistics give no strong indication one way or the other as to whether the deaths are more or less likely to be SIDS than homicides’

From: Hill, R. (2004) Multiple sudden infant deaths – coincidence or beyond coincidence, *Pediatric and Perinatal Epidemiology*, **18**, 320-326 ([http://www.cse.salford.ac.uk/staff/RHill/ppe\\_5601.pdf](http://www.cse.salford.ac.uk/staff/RHill/ppe_5601.pdf))

# Sampling strategies

- Stratified random sampling
  - improves representativeness of sampling when homogeneous sub-groups exist i.e. population is not continuous
  - Divide a population into homogeneous subpopulations (strata) and sample independently.
- Strata should be mutually exclusive: every element in the population must be assigned to only one stratum.
  - E.g. voting intentions – not a continuous variable
  - Deliberately sample groups which might be missed in a random sample e.g. small ethnic groupings

# Sampling strategies

- Various strategies for stratified random sampling
- E.g. i) Proportionate allocation
  - sampling fraction in each strata proportional to total population e.g. for 60% in the male stratum and 40% in the female stratum, then the relative size of the two samples (three males, two females) should reflect this proportion
- E.g. ii) Optimum/disproportionate allocation
  - more samples taken in strata with the greatest variability
  - E.g. if variance of women's height twice that of men, sample twice as many women as men

# Sampling strategies

- Useful for all kinds of spatial, temporal measurements
  - Stratify according to population density for e.g. to overcome density disparity
  - E.g. random samples of population in UK will lead to large bias towards SE & few/no samples in N/NE
  - Stratify according to population e.g. deliberately select areas in NE to avoid bias cause by population of SE

# Summary

- Consider sources of error (random, systematic)
- Consider best experimental design to minimise error: sampling strategy, sample size etc.
- Include some uncertainty analysis
  - at very least, quote results of sampling with some estimate of standard error



## Reading

### Various texts

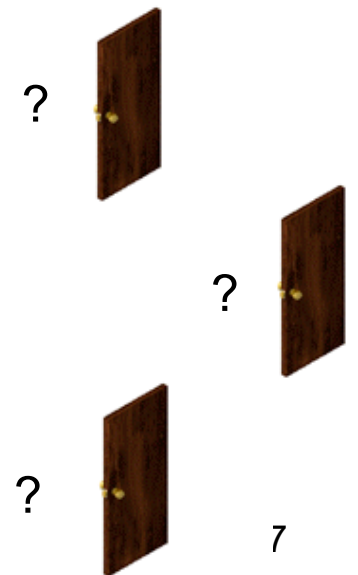
- Hardisty, J et al., *Computerised Environmental Modelling: A Practical Introduction Using Excel (Principles and Techniques in the Environmental Sciences)*, 1993, Wiley Blackwell.
  - Wainwright, J. and Mulligan, M. (eds) *Environmental Modelling: Finding Simplicity in Complexity*, 2004, John Wiley and Sons.
  - Casti, John L., 1997 *Would-be Worlds* (New York: Wiley and Sons).
- Advanced texts
- Gershenfeld, N. , 2002, *The Nature of Mathematical Modelling*, CUP.
  - Boeker, E. and van Grondelle, R., *Environmental Science, Physical Principles and Applications*, Wiley.
  - Gauch, H., 2002, *Scientific Method in Practice*, CUP.
  - Flake, W. G., 2000, *Computational Beauty of Nature*, MIT Press.

# Common errors: reversed conditional

- After Stewart (1996) & Gauch (2003: 212):
  - Boy? Girl? Assume  $P(B) = P(G) = 0.5$  and independent
  - For a family with 2 children, what is  $P$  that other is a girl, given that one is a girl?
- 4 possible combinations, each  $P(0.25)$ : BB, BG, GB, GG
- Can't be BB, and in only 1 of 3 remaining is GG possible
- **So  $P(B):P(G)$  now 2:1**
  - Using Bayes' Theorem:  $X = \text{at least 1 G}, Y = \text{GG}$
  - $P(X) = \frac{3}{4}$  and so  $P(X \cap Y)/P(X) = \frac{1/4}{3/4} = 1/3$

# Probability is a funny thing

- The Monty Hall “Paradox”
- 3 doors, behind one is a prize (Monty knows which one)
- I choose a door. Monty then opens one of the other doors without a prize and asks me if I want to change my choice
- Should I change? Does it make any difference?



# Monty Hall redux

- You should always change. But why – surely the odds are 50:50?
  - Think about possible range of outcomes:
    - Pick right door to start: 1 in 3 chance
      - Both remaining doors blank so changing after Monty opens a blank door means we always lose
    - Pick wrong door to start: 2 in 3 chance
      - Remaining doors are 1 blank and 1 with prize so Monty must open only blank door left – changing now means we always win
    - So if we always change we win  $2/3$  of the time, if we don't we only win  $1/3$  of the time
  - Can be stated as a Bayesian conditional problem

# Latin hypercube

- Used to sample N-dimensional space very sparsely
  - a square grid containing sample positions is a Latin square if (and only if) there is only one sample in each row and each column. A Latin hypercube is the generalisation of this concept to an arbitrary number of dimensions
  - Each variable in a system is guaranteed to be sampled once in each dimension (not so for random sampling)

# Latin hypercube

- E.g. I have two variables A and B I want to sample with respect to each other

A

	X			
	X	X		
B	X			

Random sampling

A

	X			
				X
			X	
B		X		

Latin square