Using statistics to improve environmental toxicology testing

A. John Bailer

Dept. of Statistics, Miami Univ., Oxford, OH 45056, USA

2012 Prob/Stat Day at UMBC

April 20-21, 2012

Acknowledgements

Thanks to Bimal Sinha for the invitation to join you!

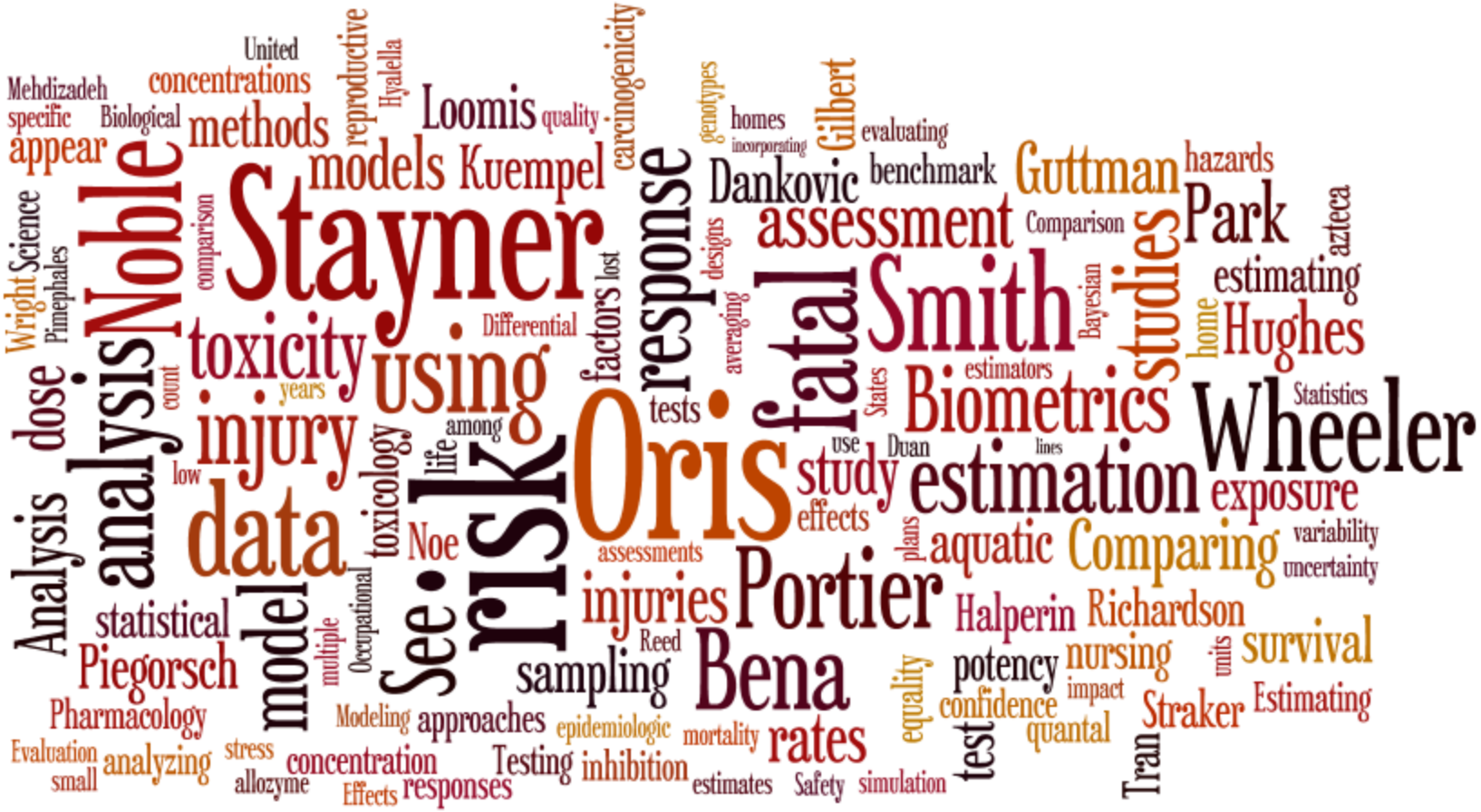
Collaborators (f.p.=first paper; m.r.p.=most recent paper):

- Chris Portier (f.p.: 1987)
- Walt Piegorsch (f.p.: 1989; f.b.: 1997)
- **Jim Oris (f.p.: 1993; m.r.p.: 2012)***
- Kyoungah See (f.p.: 1998)
- Mike Hughes (f.p.: 2000)
- Matt Wheeler (f.p.: 2003; m.r.p.: 2012)*[former student]
- Bob Noble (f.p.: 2005)
- Steve Wright (f.p.: 2006)
- Jing Zhang (f.p.: 2012)

Students:

- S. Fore' + M. Schlueter + L. Barghusen + S. Gilbert + Y. Duan + S. Walker + K. Venis + S. Liu + M. Smith + R. Elmore + J. Shumate + J. Bena + E. Cho + S. Greven + B. Bell + J. Craft + W. Fadel + S. Anderson + B. Sigal

[Note: collaborators addressing env./eco. issues – omitted occupational health + gerontological applications]



http://www.wordle.net/ (using journal articles – removing journal titles)

Outline

1. Overview
2. Potency estimation
3. Step-stress study analysis
4. Optimal design for nonlinear experiments
5. What's next?

1. Overview

Environmental toxicology = study of toxicants expressed in any environment and at any level of biological organization (from molecular to ecosystem level)

As a special case, aquatic toxicology experiments are conducted to evaluate the potential impact of chemicals in receiving waters, marine systems, and other ecosystems.

Who cares about environmental toxicology?

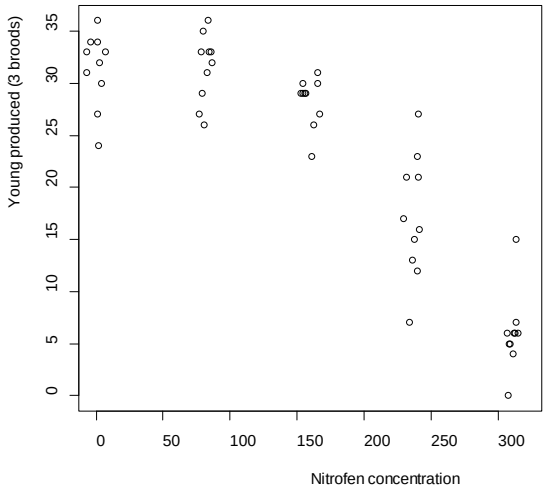
- Manufacturers of fertilizers and pesticides
- Consumer products (e.g. laundry detergents)
- Pharmaceutical industry (e.g. endocrine active compounds)
- Oil production companies (e.g. spills, Deepwater Horizon)
- We do.

2. Potency Estimation

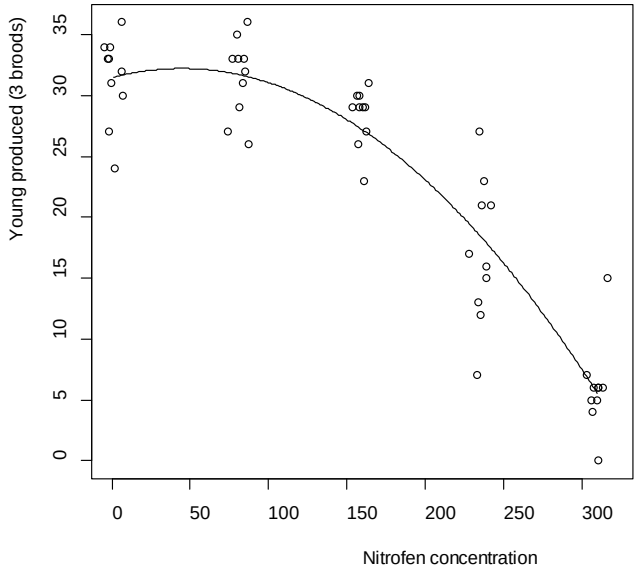
- Start of my collaborations on environmental / eco / aquatic toxicology problems.
- Question: How are regulatory limits set for exposure limits? Here the issue was targets for whole effluent toxicity testing.
- Observation: many advocated the use of NOEC/LOEC (so-called “hypothesis testing” approaches in the literature)
- Statisticians had argued for regression based alternatives (e.g. BMD as suggested by Crump 1984 although goes back to Bliss in the 1930s for LC estimation)
- Colleague wonders about sensitivity to detect differences in common protocols and if better alternative exists.

Process:

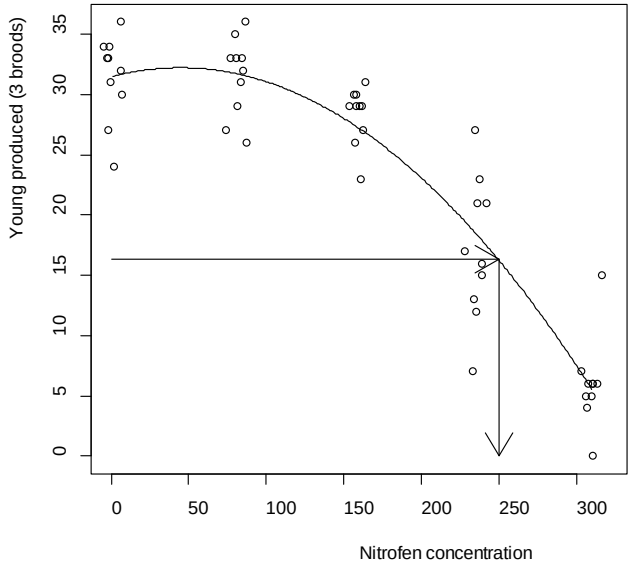
- Conduct D-R experiment
  - e.g.:
    - response (number of young produced in 3 broods)
    - dose (effluent conc. Or toxicant level)



- Fit model
  - e.g.:
    - Poisson regression



- Invert fitted D-R relationship



Concentration (dose) associated with a specific level of inhibition relative to control results, are often estimated

Test system:

Freshwater quality test based on young produced by *Ceriodaphnia dubia*:

http://denr.sd.gov/des/sw/wet.aspx

Test is typically run for seven days or less. Measures whether the discharge effects the reproduction of the Ceriodaphnia. Chronic test also measures if the effluent would be lethal to aquatic life (from same web site).

Frequentist approach [Bailer and Oris (1993, 1997)]

$$Y_{ij} \mid \mu_i \overset{independent}{\sim} Poisson(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 c_i + \beta_2 c_i^2 + \ldots + \beta_m c_i^m$$

Resulting in a potency estimate RIp: C s.t.

$$\mu_{RIp} = (1 - p)\mu_0$$

Notes:

1) m = 2 often works
2) “p” often 0.25 or 0.50
2) CI based on parametric bootstrap

Bayesian approach [Zhang, Bailer, Oris (2012)]

Parameters: $\theta = (\beta_0, \beta_1, \beta_2)$

Data: $Y = (Y_{01}, \ldots, Y_{0n[0]}, \ldots, Y_{g1}, \ldots Y_{gn[g]})$

$$f(\theta \mid Y) = \frac{f(\theta) f(Y \mid \theta)}{f(Y)}$$

where $f(Y \mid \theta)$ is the Poisson likelihood from above

$f(\theta)$ is the prior distribution for the parameters

$f(Y)$is a normalizing constant

Priors:

$$\beta_i \sim N(\beta_i^0, \sigma_i^2)$$

$$\sigma_i^2 \sim Inv-Gamma(0.001, 0.001)$$

Comment:

Historical information can be used to modify the priors or as part a hierarchical specification of an analysis.

Bayesian models	Prior distribution for β_0	Parameters used in the prior distribution	
		β_0^0	σ^2
Informative prior	$N(\beta_0^0,\sigma^2)$	3.4	0.001
Flat prior, centered at 0	$N(\beta_0^0,\sigma^2)$	0	σ^2~Inv-Gamma(0.001,0.001)
Flat prior, centered at ln(20)	$N(\beta_0^0,\sigma^2)$	ln(20)	σ^2~Inv-Gamma(0.001,0.001)
Flat prior, centered at ln(30)	$N(\beta_0^0,\sigma^2)$	ln(30)	σ^2~Inv-Gamma(0.001,0.001)

Methods	RI_{25} (true value: 0.86)				
	Frequentist Method	Bayesian Methods With flat priors for β_1 and β_2 and different priors for β_0			
		Flat prior, centered at 0	Flat prior, centered at ln(20)	Flat prior, centered at ln(30)	Informative prior
Avg. Point Est.	0.85	0.82	0.82	0.82	0.82
Root MSE	0.09	0.08	0.08	0.07	0.06
Observed Coverage of nominal 95% interval	73.30%	94.50%	94.70%	95.70%	96.30%

3. Step-Stress Test [or *is there a better way to test for differences*]

- used where the water velocity is increased incrementally over time until the fish can no longer maintain position in a chamber

- time until fatigue measured recorded

- critical swimming speed calculated (Ucrit) [a measure related to the water velocity at the time of failure]

- analysis of continuous response common (e.g. t-tests, anova…) (Kolok 1999)

- doesn't take into account structure of time-to-event data … possible loss of statistical power?

Goals:

- COMPARE groups with respect to survival observed in a step-stress study (with covariates also possible)
- Employ a parametric survival analysis approach to step-stress data
- Contrast method with continuous response alternatives

Notation:

- n = number of fish on test
- $\tau_0 = 0$ [time start]
- R_1 = stress in the first interval, water velocity
- τ_j=fixed time when stress/water velocity is increased to R_{j+1},
- $j = 1, \ldots, k\text{-}1$ [stress assumed constant within interval]
- $\tau_k = \infty$ [i.e. failure times are observed]
- $I_j = [\tau_{j-1}, \tau_j)$ represent the jth interval or step, $j = 1, \ldots, k$

Specification and conduct of the study

- number of steps (k)
- spacing of times ($\tau_0, \ldots, \tau_k$)
- water velocity at each step (R_j)
- Each fish starts swimming at $\tau_0 = 0$ and failure time t is recorded

Analysis goal

- model distribution of the failure times
- compare different groups w.r.t. failure time distributions
- quantify covariates influence on the failure time distribution.
- all failures observed (common in tox. testing apps; however, not serious problem if censoring occurs)

Statistical Model:

- piecewise constant hazard function in each interval
- h(t)= $\lambda_j \qquad t \in I_j, \lambda_j > 0, j = 1, \ldots k$

$$S(t) = \exp\{-\sum_{l=1}^{j-1} \lambda_l (\tau_l - \tau_{l-1}) - \lambda_j (t - \tau_{j-1})\}$$

$$L = (\prod_{j=1}^{k} \lambda_j^{N_j}) \exp(-\sum_{j=1}^{k} \lambda_j T_j)$$

Refs:

Greven S., Bailer A.J., Kupper L.L., Muller K.E. and **Craft, J.L.**(2004)
Craft J.L. and Bailer A.J. (2005)

Test for a treatment effect - H_0: $\alpha_j = 0 \; \forall \; j = 1, \ldots, k$
(identical hazard/survival)

$$-2\ln\hat{\lambda} = 2\sum_{j=1}^{k}\{N_{0j}\ln(\frac{N_{0j}(T_{0j}+T_{1j})}{T_{0j}(N_{0j}+N_{1j})}) + N_{1j}\ln(\frac{N_{1j}(T_{0j}+T_{1j})}{T_{1j}(N_{0j}+N_{1j})})\}$$

Example:

- Fish were swum at 15 cm/s for 90 minutes
- flow rate increased every 20 minutes by 5 cm/s until fatigue.
- Control (n=14; mean = 118.3, median = 114.7, s = 23.4)
- Treatment (n=15; mean = 129.8, median = 117.3, s = 32.6)
- Combining adjoining intervals b/c small n_i - [0, 110), [110, 150) and [150, ∞) min. - water velocities of 15-20 cm/s, 25-30 cm/s and 35-40 cm/s, respectively.

Comments:

- If the time interval lengths are chosen too short in relation to the water velocity and the attendant hazard rates, there may be no observed failures in some of the intervals
- Design issue? specify interval lengths that are roughly inversely related to water velocity. This may serve as a surrogate for specifying interval lengths that are inversely related to the hazard of failure in each interval?

Context for design investigation

- Testing for ecological impacts of toxins can occur in lab / field
- methyl tert-butyl ether (MTBE) is an oxygenate used in reformulated gasoline [decrease CO emission with fossil fuels].
- Fluoranthene in exhaust of internal combustion engines
- Water craft such as two-cycle jet skis release 20-30% of fuel is in exhaust unburned.
- Implication: unburned fuel entering aquatic systems

Q: does this exposure impact organisms in aquatic systems?

Q: mix of fluoranthene+MTBE worse than fluoranthene alone?

Test system: Fathead minnow larvae

(image from www.eeusa.com/?cat=2)

Endpoints modeled: fluoranthene body residue (FL vs. FL+MTBE) and survival

Experiment:

Time-course for body residue + survival-time study

Model:

Toxicokinetic model involving uptake and elimination (more to come) + survival (factorial treatment structure with FL & MTBE)

Result? [Cho, Bailer and Oris 2003 *Environ. Sci. Technol.* **37**:1306-1310]

- Higher FL body residue when MTBE present
- Lower survival with co-exposure [MTBE enhances photo-induced toxicity of FL]

Spin-off question … are we sampling at the best possible times …

- Statistical design concepts provide direction for the spacing and allocation of experimental organisms for optimal estimation of system characteristics

- 4. Optimal design for nonlinear expts.

- **Details**: experiment with 72 h of contaminant exposure followed by 72 h with no contaminant exposure.

- Internal concentration of toxin in the organism ↑ during exposure and ↓ after exposure ends.

- Some sampling times fixed *a priori* - 0h=start, 72h =cessation of exposure, and 144h=end of study.

- Resources available to support the selection of 10 additional sampling times.

- So, “design” = particular set of 10 unique sampling times (besides) 3 fixed times.

Assume:

- sampling on the hour is desired (e.g., for logistical reasons such as scheduling technicians) → 145 possible sampling times (from 0h through 144h);
- three of times (0, 72, and 144h) are designated by the practitioners, and ten are to be selected from the remaining times.

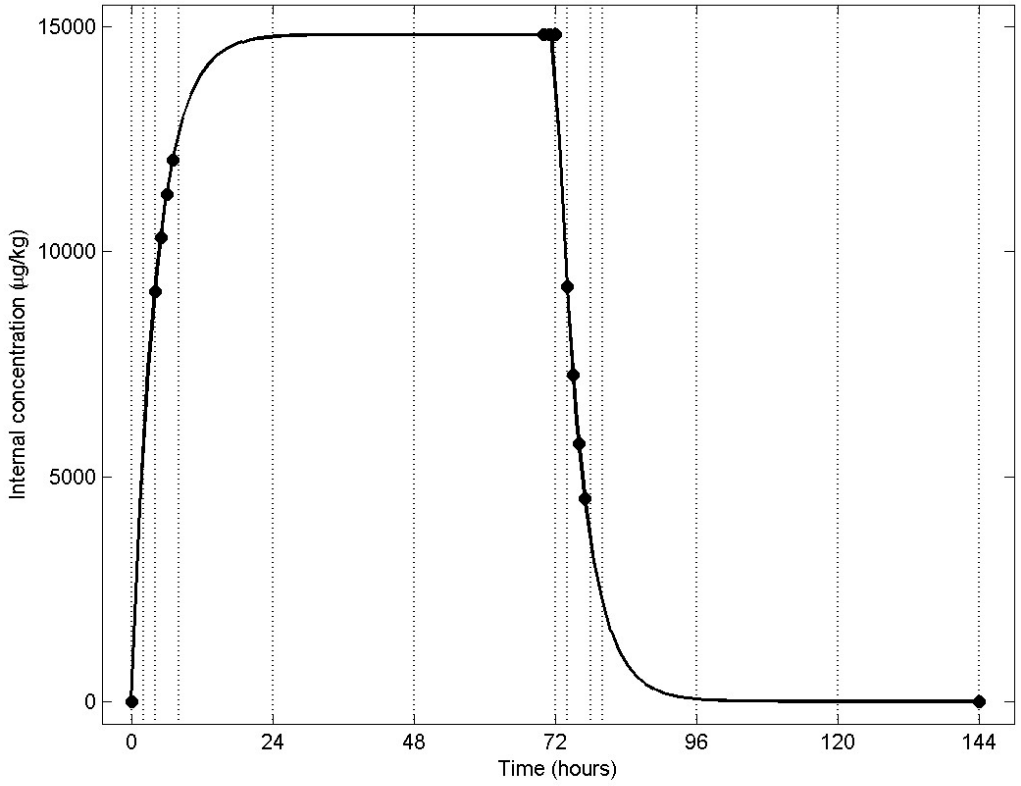
"**design**" = particular set of 10 unique sampling times (besides the three fixed times) - # possible designs = chose 10 elements from a set of 142 elements = 6.6×10^{14}.

Goal (restated): develop strategies to determine which among the 664 trillion possible designs should be preferred, according to some combination of fixed criteria.

Mean internal concentration at time *x*

[reflects the uptake and elimination of the toxin]

$$\mu(x;\theta_1,\theta_2) = \frac{\theta_1}{\theta_2}\left[\exp(-\theta_2 \cdot \max\{x-72,0\}) - \exp(-\theta_2 \cdot x)\right], \qquad (1)$$



D-optimality criterion

Select design points from the integers to maximize the function (of inform. Matrix)

* D-*optimality criterion*, where D = **det** M(x_1,…x_{13})

[Common design criterion (Atkinson and Donev 1992)]

$$\mathbf{M} := \sum_{x \in S} \left[J_{\theta} F(x; \hat{\theta}) \right]^T \left[J_{\theta} F(x; \hat{\theta}) \right]$$

$J_{\theta} F(x; \hat{\theta})$ = Jacobian (partial derivatives) of the mean response function w.r.t. the parameters, $\hat{\theta}$ given *a priori* estimate of θ

S = experimental design (i.e., a set of sampling points chosen for the experiment).

* Choose S so that $S_0 \subseteq S \subseteq X$

  X = the finite design space of all possible sampling points

  S_0 = a set of required design points

* Optimal design is an example of a *nonlinear knapsack problems* (Bretthauer and Shetty 2002): **maximize**

$f(S) = \varphi\left(\sum_{x \in S} \mathbf{M}_x\right)$ over all sets S subject to $S_0 \subseteq S \subseteq X$ and

$$c(S) = \sum_{x \in S} \gamma_x \leq \beta.$$

γ_x = cost for sampling point x, and β denotes the total available budget. D-optimality criterion: use $f(S) = D(S) = \det \sum_{x \in S} \mathbf{M}_x$.

Equal costs (γ_x =1 for all $x \in X$):

knapsack constraint

$$\sum_{x \in S} \gamma_x \leq \beta$$

becomes a **cardinality** constraint of form $|S| \leq k$ in this case

Greedy Method (cardinality-constrained form)

1. Initialize $S := S_0$.
2. While $|S| < k$ do the following:
   a. Choose any $t \in X \setminus S$ for which $f(S \cup \{t\})$ is maximized;
   b. Replace S by $S \cup \{t\}$.

backslash symbol '\' = subset exclusion, so that

$A \setminus B$ = {all members of set A that are not members of set B}

* Step 2a above:

t is selected from points not already in the design S

Iterative Replacement Method (cardinality-constrained form)

1. Choose an initial subset s satisfying $|S| = k$ and $S_0 \subseteq S \subseteq X$.
2. Take $\overline{S} := S$ and $T := \overline{S} \setminus S_0$.
3. While T is nonempty, do the following:
   a. Choose $t \in T$ and replace T by $T \setminus \{t\}$;
   b. Choose any $r \in X \setminus S$ for which $f(\{r\} \cup (S \setminus \{t\}))$ is maximized;
   c. If $f(\{r\} \cup (S \setminus \{t\})) > f(S)$, then
      replace S by $\{r\} \cup (S \setminus \{t\})$.
4. If $S = \overline{S}$, then stop; otherwise, go to step 2.

Step 1: initial subset *S* might naturally be found using greedy method.

Step 2: temporary set *T* consists of all non-required members of current set *S*.

Step 3: checks each member of *T* in its search for a fruitful exchange of design points.

Step 3c: requirement of strict increase + finiteness of the search space *X*, guarantees that iterative replacement method eventually terminates.

Step 4: stops algorithm if step 3 has failed to improve set *S*.

Notes:

* experimental design literature, cardinality-constrained greedy method is known as “sequential search” (Dykstra 1971)

* cardinality-constrained iterative replacement is known as “modified Fedorov” method (Cook and Nachtsheim 1980)

* combination of cardinality-constrained greedy and iterative replacement methods provides a two-phase procedure shown to be among the most competitive for a wide variety of linear models (Atkinson and Donev 1992) [adopted by the SAS Institute for the OptEx procedure of the SAS/QC module]

Wright and Bailer (2006) [*Biometrics* **62**: 886-892] used the procedure in the context of the fluoranthene experiment under the assumption of uniform sampling costs.

Result of optimization, compared with practitioner's original design, was as follows:

- Optimized design — {0h,4h,5h,6h,7h,70h,71h,72h,74h,75h,76h,77h,144h};

- Practitioner's design — {0h,2h,4h,8h,24h,48h,72h,74h,78h,80h,96h,120h,144h}.

Optimized design: $D(S) = 3.81\times10^{11}$

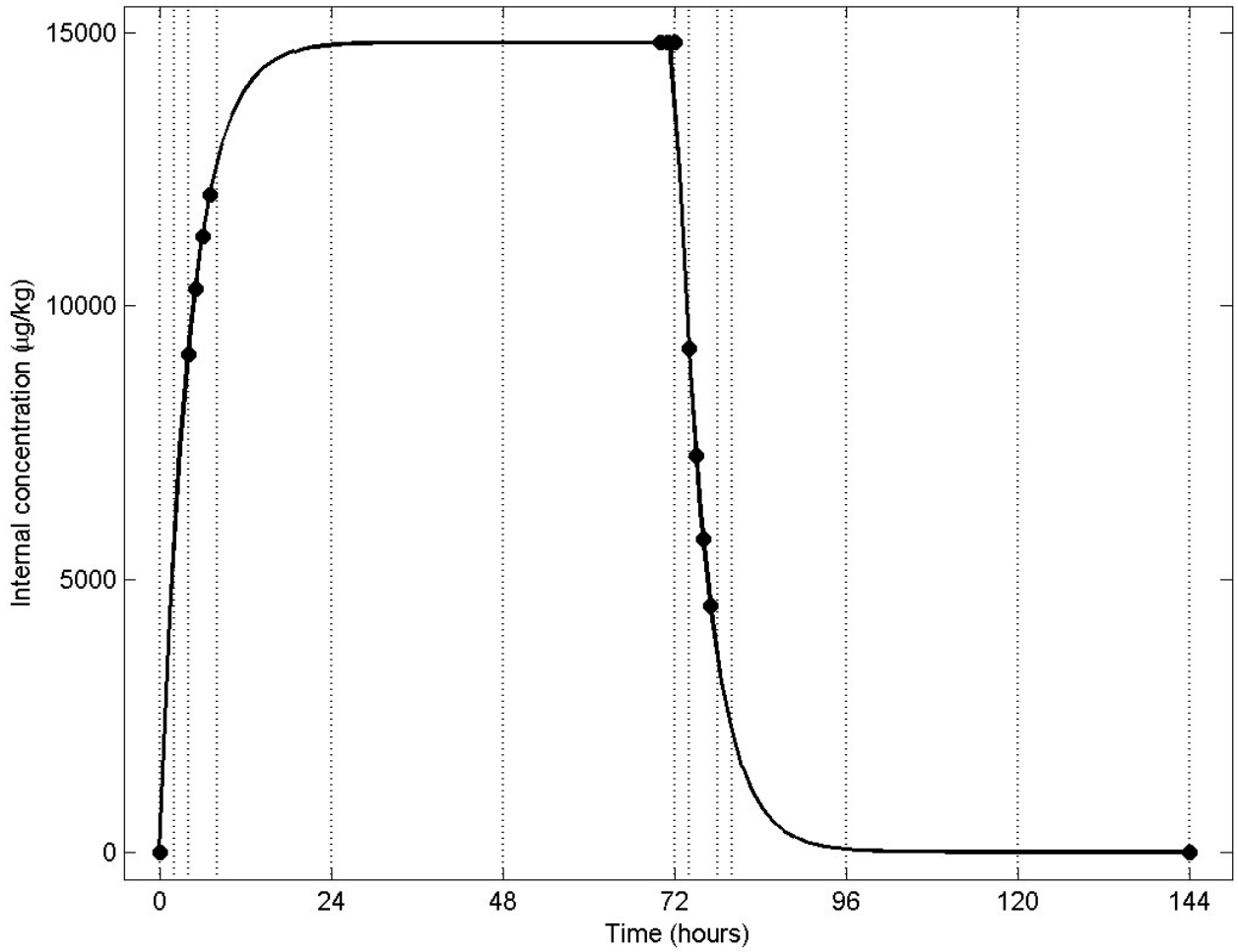
Practitioner's design: $D(S) = 1.84\times10^{11}$

Comparing the sampling times in these two designs

* practitioner correctly chose to select sampling times:
  > near start of study (times before 10h)
  > near time when exposure ceases (times between 70h and 78h).

* Practitioner's design includes five sampling times deemed unimportant by the optimal design, but which may still be important for non-statistical reasons.

Figure 1. Response function (uptake and elimination of toxin) shown with practitioner's sampling times (vertical dotted lines) and the **optimized experimental design** (**large dots**) assuming uniform sampling costs



Comparing different number of design points

[D(*) scaled some differently than other displays but still reflect the actual relative magnitudes among designs.]

Table 2: Designs Obtained for Different Numbers of Samples

	Sampling times appearing in design (h)													
No.	4	5	6	7	8	9	70	71	73	74	75	76	77	$D(\cdot)$
9		*	*	*						*	*	*		41.37
10		*	*	*				*		*	*	*		53.57
11	*	*	*	*				*		*	*	*		66.73
12	*	*	*	*				*	*	*	*	*		81.41
13	*	*	*	*			*	*		*	*	*	*	98.17
14	*	*	*	*	*		*	*		*	*	*	*	115.58
15	*	*	*	*	*		*	*	*	*	*	*	*	134.39
16	*	*	*	*	*	*	*	*	*	*	*	*	*	153.54

NOTE: The times 0, 72, and 144 h were required in the design.

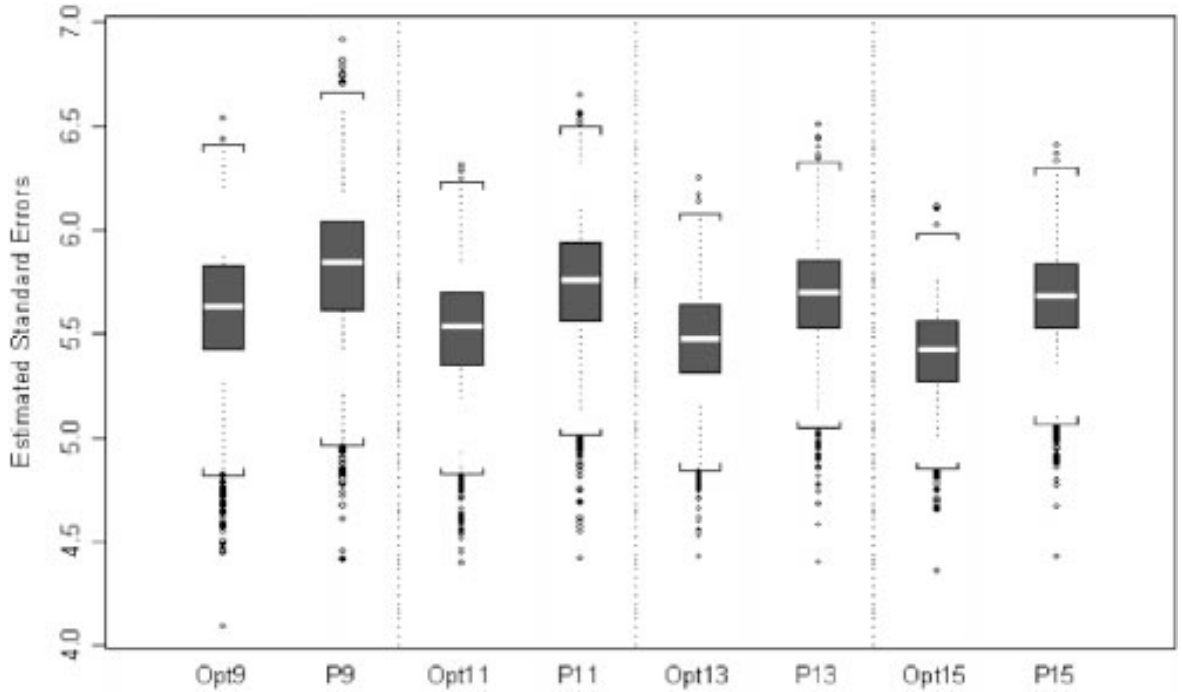
Results – sensitivity analysis - 10000 samples:

- θ_1/ θ_2 ~ triangular
- θ_2~ uniform with extremes set at ±10% best value

	Sampling times appearing in designs(h)												
Profile	4	5	6	7	8	70	71	73	74	75	76	77	Freq.
1	*	*	*	*	*		*	*	*	*	*		4982
2		*	*	*	*	*	*		*	*	*	*	2559
3	*	*	*	*		*	*		*	*	*	*	2459

Simulation Study [Bell, Bailer and Wright (2006) *ET&C* **25**:248-252]

- $\theta_1 = K_u$ = 3532 and $\theta_2 = K_e$ =0.2358 { *uptake/elimination lingo* }
- Case 1: constant variance / normal
- Case 2: proportional variance / log-normal
- nonlinear regression model fit (SEs for K_u estimation)



To recap …

- At this point we have addressed – WHEN, WHERE, and HOW LONG

Next steps ... HOW MUCH

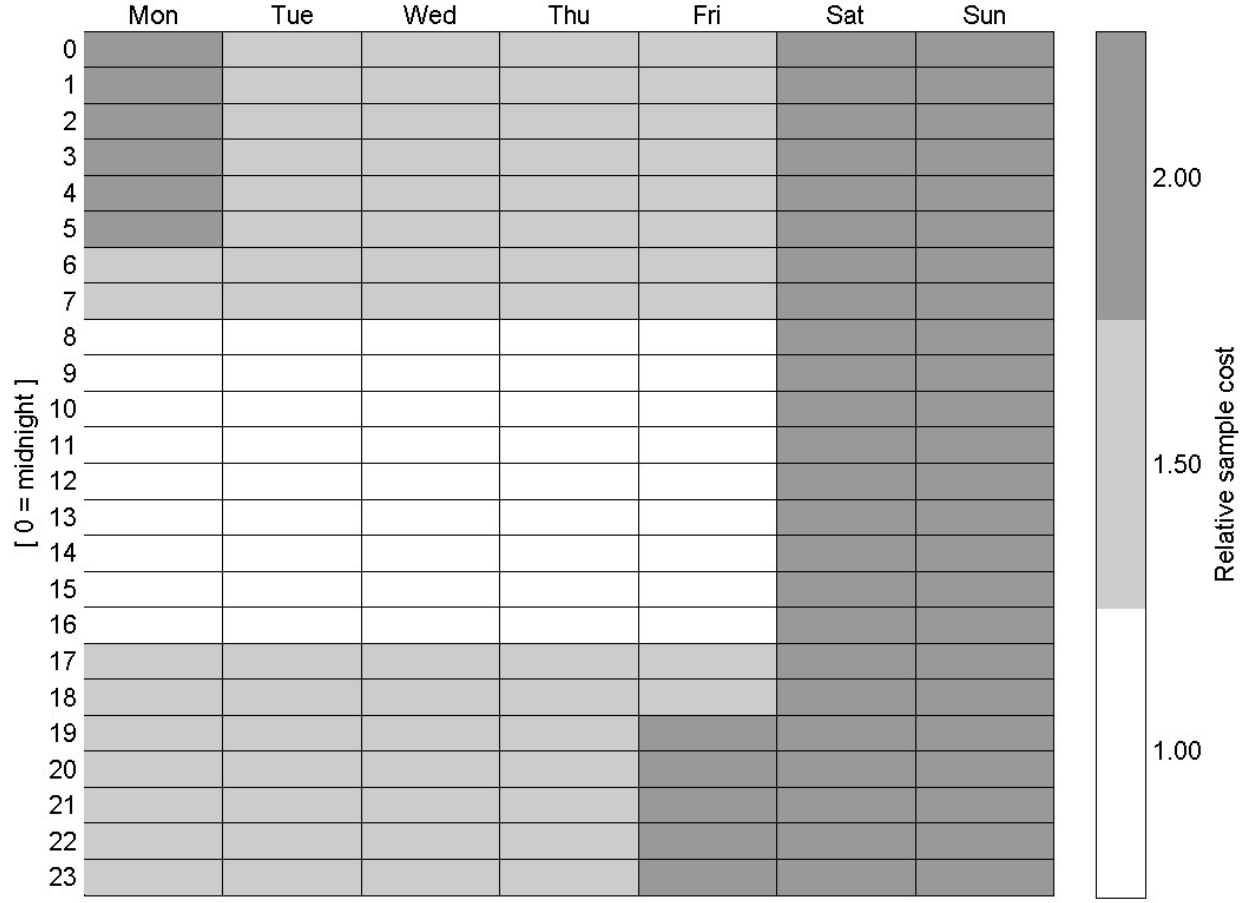
- Constrain choice of designs to accommodate practical considerations.
- Motivation: conducting studies on weekdays may be cheaper than weekend/evening (overtime pay, etc.) → determine best design when sample times are constrained to occur between 8 a.m. and 5 p.m. on weekdays.
- Cost of conducting the experiment = *f*(times when observations are sampled and processed)
- Why? laboratory technicians may be paid at different hourly rates depending on the time of day or on the day of the week

- Consider relative cost of starting experiment described previously at 2 different starting times in week

Assume:

- practitioner's sampling design in both cases.
- sampling costs vary throughout week as follows (see Figure):
- $1\times p\times w$ = "base/regular hourly wage" on weekdays (8 a.m.–5 p.m., Monday through Friday);
- $2\times p\times$w = "double-time wage" on the weekend (7 p.m. Friday through 6 a.m. Monday);
- $1.5\times p\times w$ = "time-and-a-half wage" at all other times.
- p=time to collect sample; w=base wage

Figure 2. Cost of one sample obtained at given hour of the week, with 1.00 = “base cost” for samples collected during normal business hours, 1.50 = “time-and-a-half” [overnight], and 2.00 [weekends].



Does it matter when you start a study?

- Case 1: start at 8 a.m. on Monday

  all samples drawn during normal work hrs, except samples at 120h & 144h (drawn on weekend =double-time wages)

- Case 2: start at 8 a.m. on Wednesday

  five samples (72h through 96h) are drawn over weekend.

- Total costs incurred:

  $15 \times p \times w$ [Case 1: Monday start]

  $21 \times p \times w$ [Case 2: Wednesday start]

⇨ cost of study would be 40% higher if started on Wednesday than if started on Monday at same time of day.

Optimal design $\subset$ combinatorial optimization known as *nonlinear knapsack problems* (Bretthauer and Shetty 2002)

* maximizing total value of items in a knapsack subject to a constraint imposed by volume capacity of knapsack.

Mathematically, we want to maximize

$$f(S) = \varphi\left(\sum_{x \in S} \mathbf{M}_x \right) \text{over all sets } S$$

$$\text{subject to } S_0 \subseteq S \subseteq X \text{ and } c(S) = \sum_{x \in S} \gamma_x \leq \beta.$$

φ = scalar aggregate of total info $\sum_{x \in S} \mathbf{M}_x$ (summed over x in S)

γ_x = cost for sampling point x

β = total available budget

Greedy Method (general form)

1. Initialize $S := S_0$.

2. While $\beta - c(S) \geq \min_{x \in X \setminus S} \gamma_x$ do the following:

   a. Choose any $t \in X \setminus S$ with $\beta - c(S) \geq \gamma_t$ that maximizes the ratio $\dfrac{f(S \cup \{t\}) - f(S)}{\gamma_t}$;

   b. Replace S by $S \cup \{t\}$

* while-condition in step 2 asks: is there room in budget to include any more design points?
* inequality in step 2a only considers unused design points that are cheap enough to be included
* choosing the new design point to maximize the specified ratio is essentially a discrete analog of steepest ascent

Iterative Replacement Method (general form)

1. Choose initial subset S satisfying

$$\beta - c(S) < \min_{x \in X \setminus S} \gamma_x \quad \text{and} \quad S_0 \subseteq S \subseteq X .$$

2. Take $\bar{S} := S$ and $T := \bar{S} \setminus S_0$.

3. While T is nonempty, do:

   a. Choose $t \in T$ and replace T by $T \setminus \{t\}$;

   b. Use Greedy Method (general form) to maximize $f(R)$ over all $R \subseteq X$ with $c(R) \leq \beta$ and $S \setminus \{t\} \subseteq R$;

   c. If $f(R) > f(S)$, then replace S by R.

4. If $S = \bar{S}$, then stop; otherwise, go to step 2.

Results (with costs considered …)

* Fixed budget = $13 \cdot p \cdot w$.

* Top four starting times that yield the most information:

1. Tuesday 9 a.m., $D(S) = 3.81058 \times 10^{11}$;

2. Tuesday 8 a.m., $D(S) = 3.81057 \times 10^{11}$;

3. Friday 9 a.m., $D(S) = 3.76002 \times 10^{11}$;

4. Friday 8 a.m., $D(S) = 3.746144 \times 10^{11}$.

Figure 4. Information (D-optimality criterion) for experimental designs optimized to best suit specific starting times, subject to *variable* sampling costs. Lighter shading indicates starting times giving more information. Given a starting time during the week, each design was chosen to maximize the D-optimality criterion, subject to a budget allowing 13 sampling times during normal business hours (or fewer, if more expensive evening or weekend hours are used) and to the requirement that sampling times must include the start, midpoint, and end of study (i.e., 0h, 72h, and 144h).

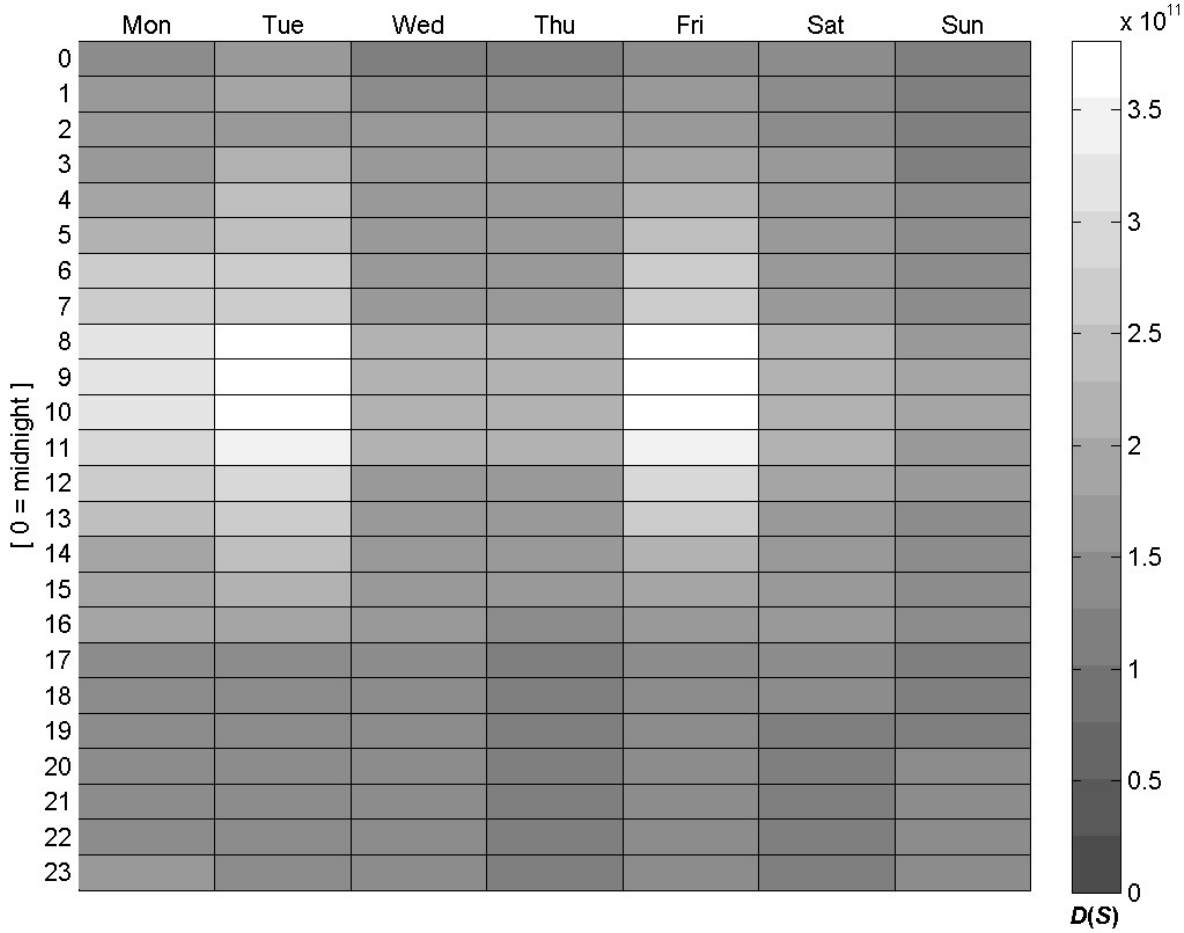
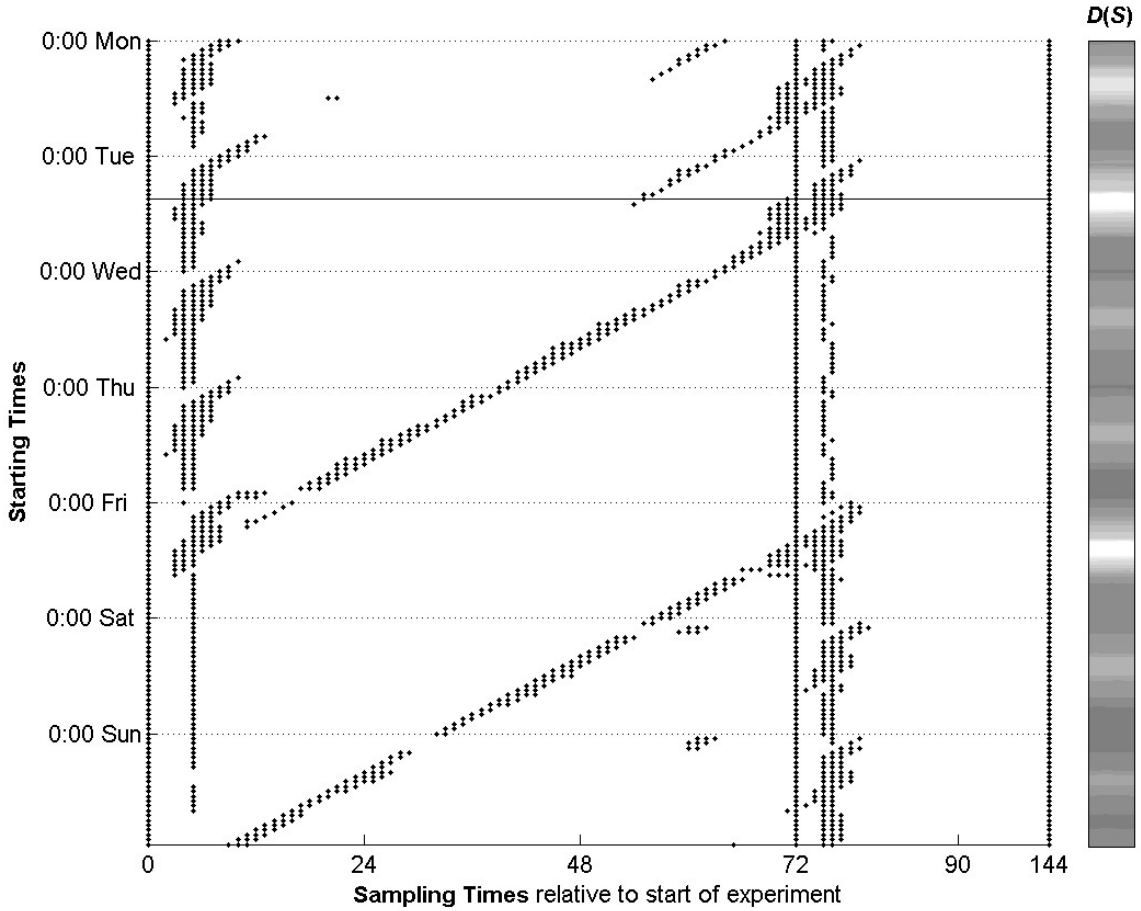


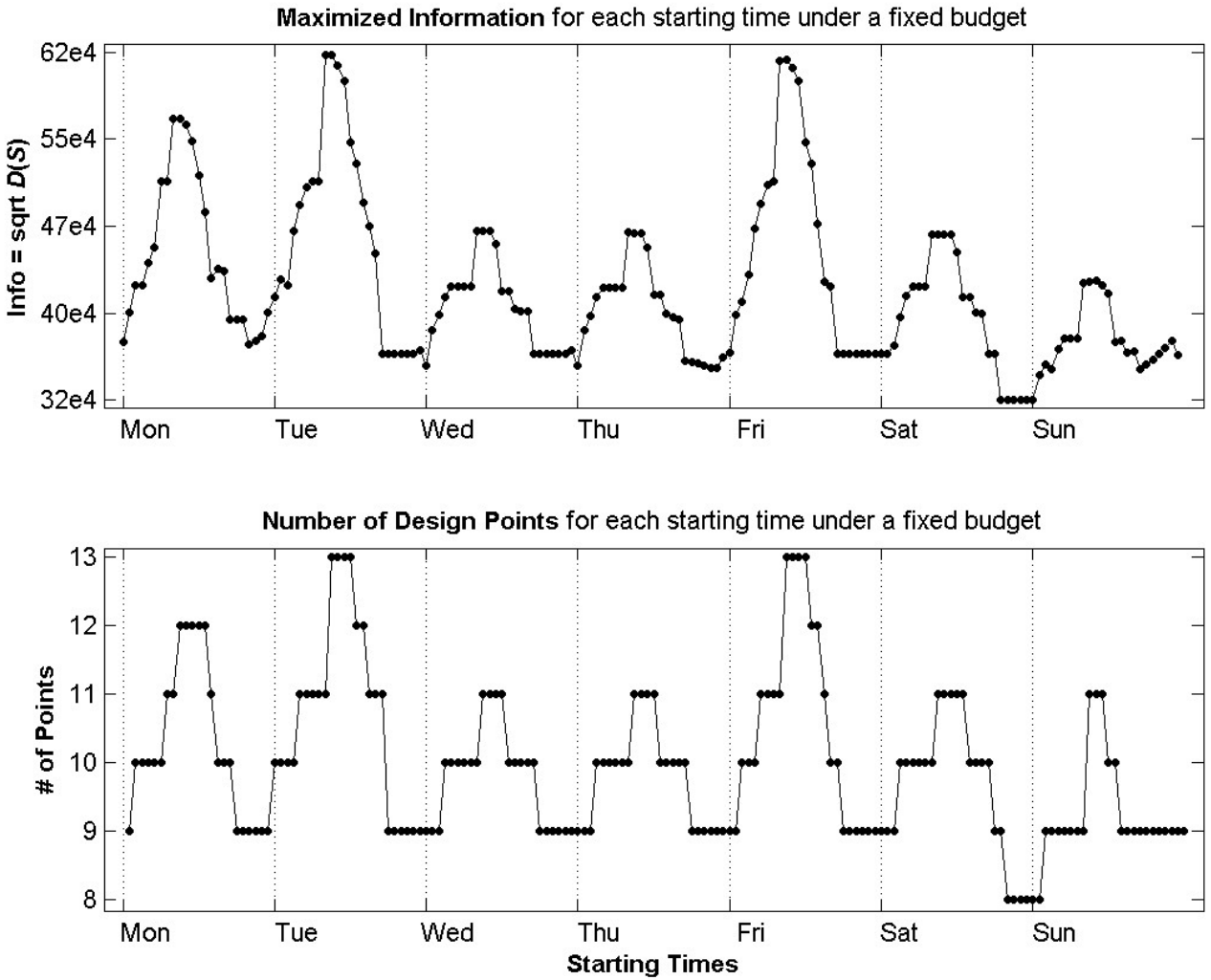
Figure 5. Optimal designs obtained for each experiment starting time under the same conditions as Figure 4. Lighter shading = designs with more information; the **solid horizontal line** =design with the **highest possible information**.



> designs with higher info. values cluster sample times ~ 5h & 72h

Relationship between maximum info and # of design points

[**Figure 6.** Information & number of design points plotted vs. designated starting time of the experiment.]



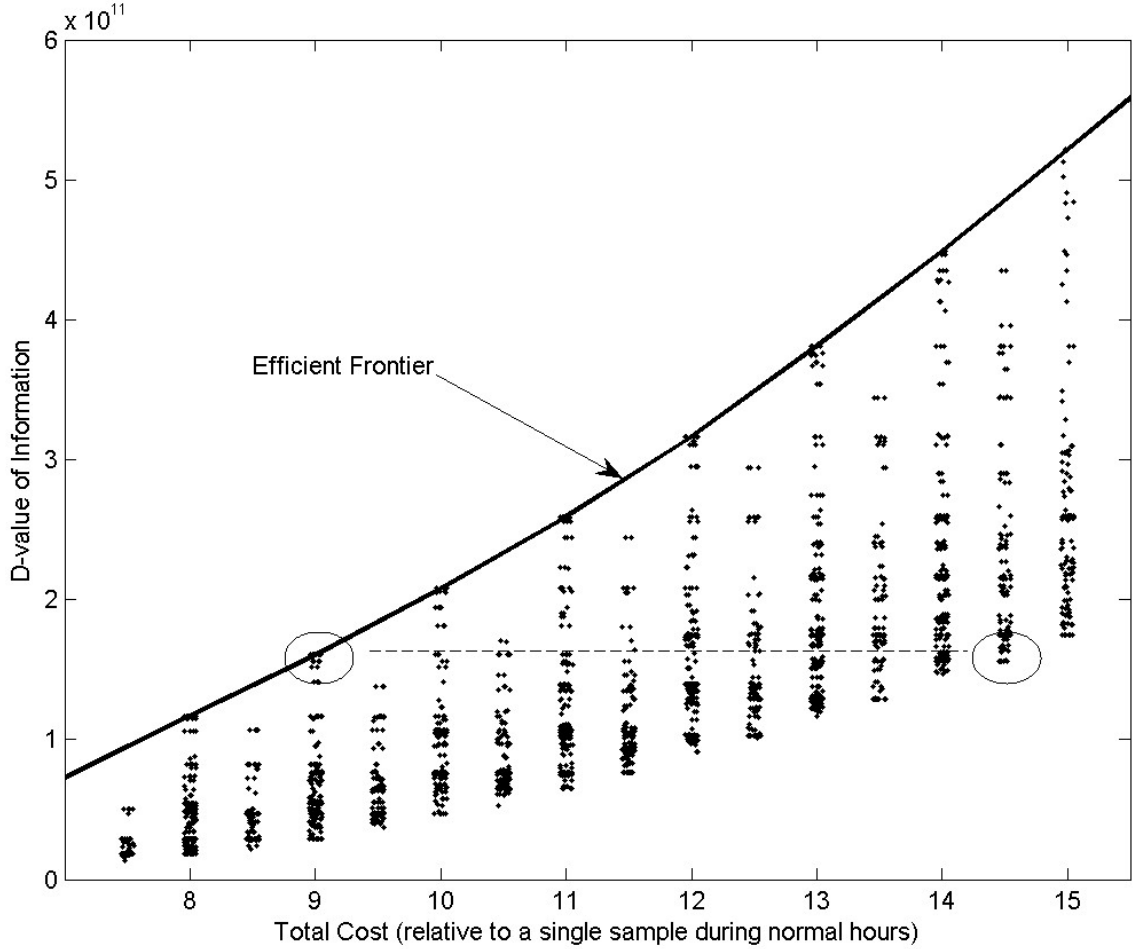
* information correlates strongly with number of samples

* desirability of samples drawn near 5h and 72h strongly favors starting the experiment near beginning of a normal workday

* number of design points in an information-optimized design drops to as few as 8 for expensive starting times.

* sampling times at start and middle of experiment are so valuable that *we would rather pay a high premium to use a few sampling times near 0h and 72h than to sample more heavily at inexpensive times!*

Trade-off between info & total cost for opt. designs - budgets ranging from 8 to 15 sampling times (at normal weekday wages). From any point on frontier, info increase only by cost increase, and cost decrease only be decreasing information



5. What's Next?

* scheduling of experiments that can overlap in time and potentially share (or prohibit sharing of) resources?

* sampling of times for design of multi-compartment PBPK/PBTK studies?

* model for hierarchical responses:

```
hatching -> survival -> growth
```

In this talk, we considered 3 problems where statistical thinking and modeling was used to

1. define and estimate a better endpoint for setting exposure limits + incorporating prior information in a potency estimation analysis

2. represent an experimental protocol in order to derive a more sensitive test

3. improve ecotoxicological designs

This work has allowed opportunities to combine ideas from regression, design, accelerated life testing to tackle problems that are important for evaluating the impact of toxicants. The chance to work closely and learn with scientific collaborators is one of the joys of statistical practice.

Thank you!