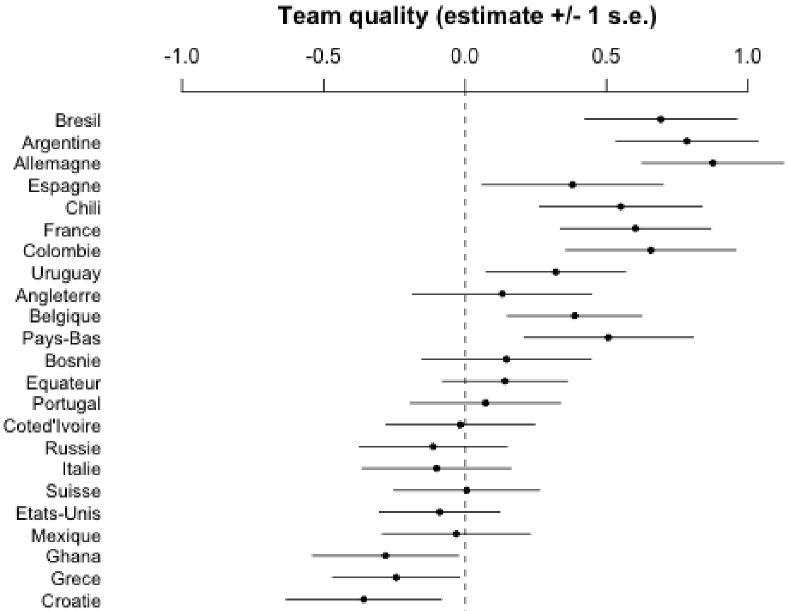
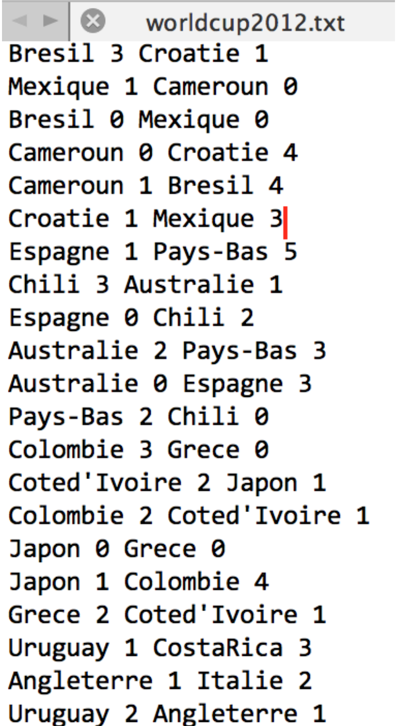
What is Bayesian data analysis? Some examples

Andrew Gelman

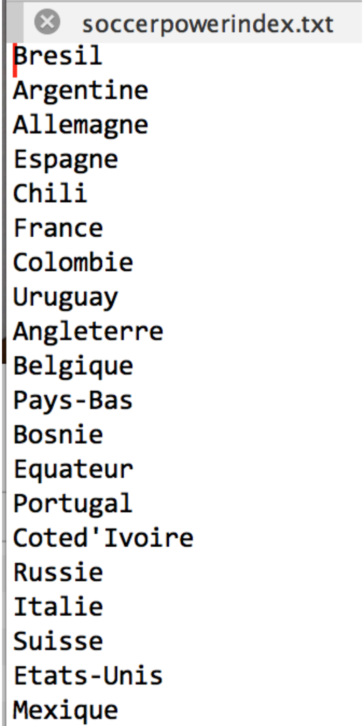
17 Feb 2016

Example 1: Stan goes to the World Cup




worldcup2012.txt

```
Bresil 3 Croatie 1
Mexique 1 Cameroun 0
Bresil 0 Mexique 0
Cameroun 0 Croatie 4
Cameroun 1 Bresil 4
Croatie 1 Mexique 3
Espagne 1 Pays-Bas 5
Chili 3 Australie 1
Espagne 0 Chili 2
Australie 2 Pays-Bas 3
Australie 0 Espagne 3
Pays-Bas 2 Chili 0
Colombie 3 Grece 0
Coted'Ivoire 2 Japon 1
Colombie 2 Coted'Ivoire 1
Japon 0 Grece 0
Japon 1 Colombie 4
Grece 2 Coted'Ivoire 1
Uruguay 1 CostaRica 3
Angleterre 1 Italie 2
Uruguay 2 Angleterre 1
```


soccerpowerindex.txt

```
Bresil
Argentine
Allemagne
Espagne
Chili
France
Colombie
Uruguay
Angleterre
Belgique
Pays-Bas
Bosnie
Equateur
Portugal
Coted'Ivoire
Russie
Italie
Suisse
Etats-Unis
Mexique
```

The model

- Fit data on signed square roots:

$$y_{ij} = \sqrt{\text{score differential when team } i \text{ plays team } j}$$

- Model $y_{ij} \sim \mathrm{N}(a_i - a_j, \sigma_y^2)$
- a_i and a_j are "ability parameters"
- σ_y is a scale parameter
- To allow for outliers, use t_7 instead of normal
- Prior info on abilities:

$$a_i \sim \mathrm{N}(\mu + b * \text{prior.score}_i, \sigma_a^2)$$

- We can set $\mu = 0$
- No further prior info in model

Stan model (part 1)

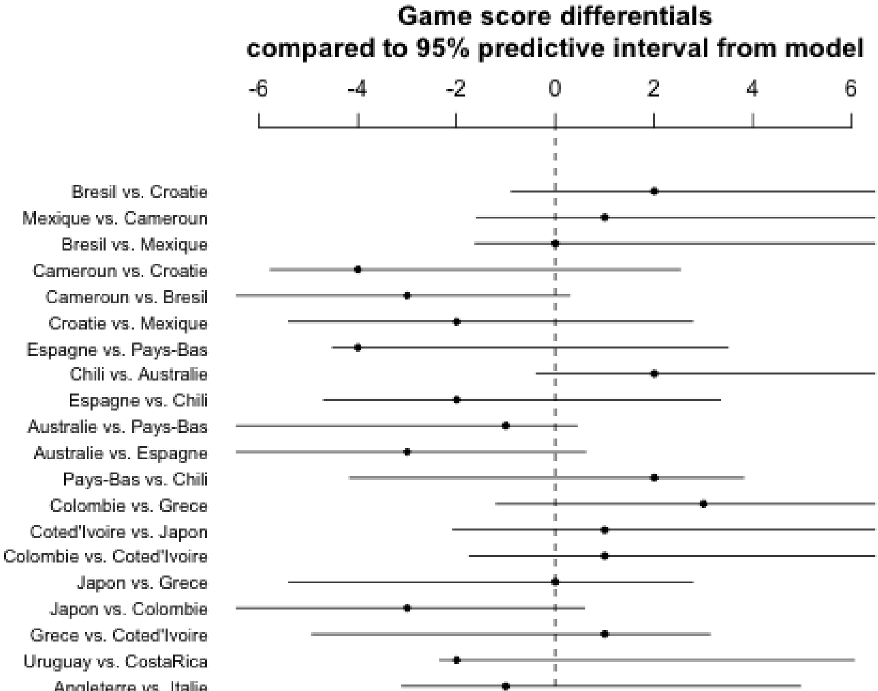
```
data {
  int nteams;
  int ngames;
  vector[nteams] prior_score;
  int team1[ngames];
  int team2[ngames];
  vector[ngames] score1;
  vector[ngames] score2;
  real df;
}
transformed data {
  vector[ngames] dif;
  vector[ngames] sqrt_dif;
  dif <- score1 - score2;
  for (i in 1:ngames)
    sqrt_dif[i] <- (step(dif[i])-.5)*sqrt(fabs(dif[i]));
}
```

Stan model (part 2)

```
parameters {
  real b;
  real<lower=0> sigma_a;
  real<lower=0> sigma_y;
  vector[nteams] eta_a;
}
transformed parameters {
  vector[nteams] a;
  a <- b*prior_score + sigma_a*eta_a;
}
model {
  eta_a ~ normal(0,1);
  for (i in 1:ngames)
    sqrt_dif[i] ~ student_t(df, a[team1[i]]-a[team2[i]],sigma_y)
}
```

Fitting the model

- Go into R
- Read in the data
- Fit the Stan model
- Check convergence
- Graph the estimated team abilities
- Re-fit without prior information
- Compare to model with prior information

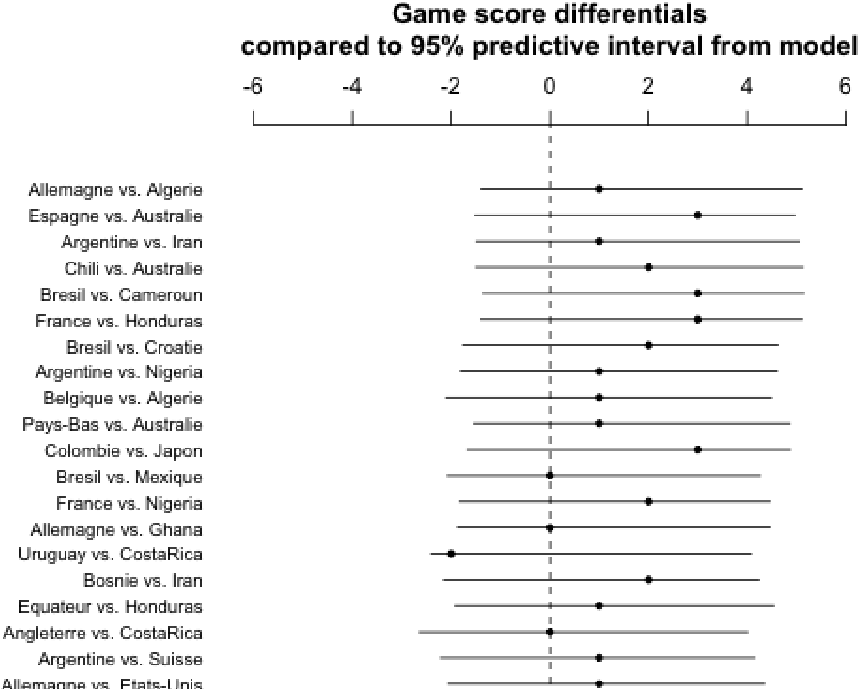

Game score differentials
compared to 95% predictive interval from model
-6
-4
-2
0
2
4
6
Bresil vs. Croatie
Mexique vs. Cameroun
Bresil vs. Mexique
Cameroun vs. Croatie
Cameroun vs. Bresil
Croatie vs. Mexique
Espagne vs. Pays-Bas
Chili vs. Australie
Espagne vs. Chili
Australie vs. Pays-Bas
Australie vs. Espagne
Pays-Bas vs. Chili
Colombie vs. Grece
Coted'Ivoire vs. Japon
Colombie vs. Coted'Ivoire
Japon vs. Grece
Japon vs. Colombie
Grece vs. Coted'Ivoire
Uruguay vs. CostaRica
Angleterre vs. Italie

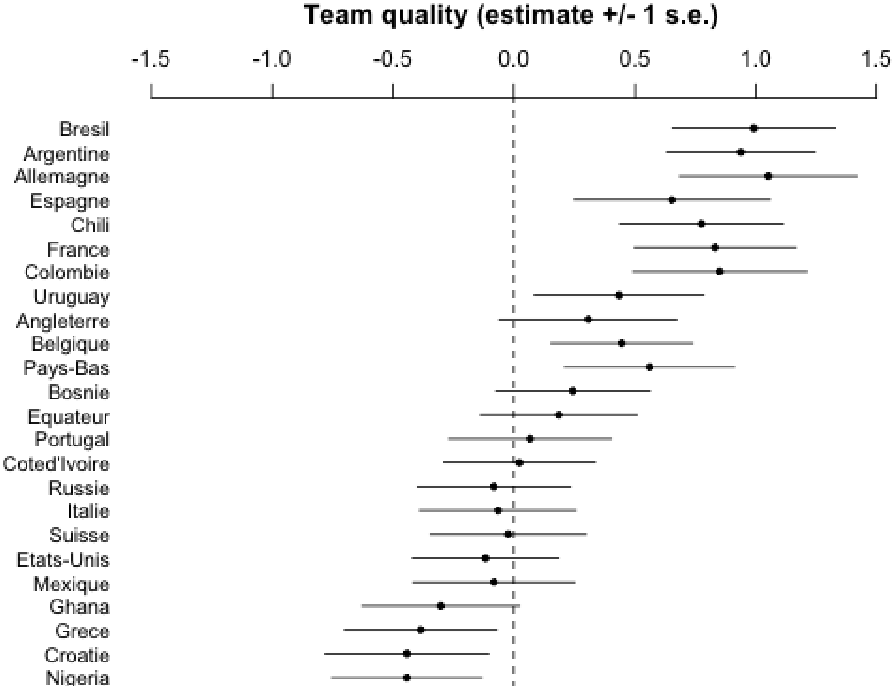
Checking model fit

- Still inside R
- For each game, plot actual score differential and 95% predictive intervals
  - Not cross-validated but no big deal in this case because *n* is large
- The predictions don't fit the data!!
- Redoing the predictive intervals
- Re-plot, still a problem!

I found the bug!

- Still inside R
- Re-fit the model on the original scale
- Display the estimated team abilities
- Updated plot of data with predictive intervals—now it's ok!
- Go back and find the bug in the square-root-scale model
- Re-fit the debugged model


Game score differentials
compared to 95% predictive interval from model
-6
-4
-2
0
2
4
6
Allemagne vs. Algerie
Espagne vs. Australie
Argentine vs. Iran
Chili vs. Australie
Bresil vs. Cameroun
France vs. Honduras
Bresil vs. Croatie
Argentine vs. Nigeria
Belgique vs. Algerie
Pays-Bas vs. Australie
Colombie vs. Japon
Bresil vs. Mexique
France vs. Nigeria
Allemagne vs. Ghana
Uruguay vs. CostaRica
Bosnie vs. Iran
Equateur vs. Honduras
Angleterre vs. CostaRica
Argentine vs. Suisse
Allemagne vs. Etats-Unis


Team quality (estimate +/- 1 s.e.)
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Bresil
Argentine
Allemagne
Espagne
Chili
France
Colombie
Uruguay
Angleterre
Belgique
Pays-Bas
Bosnie
Equateur
Portugal
Coted'Ivoire
Russie
Italie
Suisse
Etats-Unis
Mexique
Ghana
Grece
Croatie
Nigeria

Bayesian data analysis: (1) Modeling

- "Generative models"
  - Data are a realization from a (multivariate) probability distribution
  - Data vector y, probability model $p(y|\theta)$, parameter vector θ
- Prior distributions
  - In Bayes inference, the parameter vector θ is a realization from a prior distribution, $p(\theta|\phi)$
  - Vector of hyperparameters ϕ is specified or itself modeled

Bayesian data analysis: (2) Inference

- Inference is represented by a matrix of posterior simulations
  - 1000 simulations of 90 parameters: a 1000×90 matrix
- Postprocessing
  - Inference for qoi's
  - Decision analysis

Bayesian data analysis: (3) Model checking/improvement

- Do the inferences make sense?
- Are the model's predictions consistent with the data?
- *Not*: Is the model true?
- *Not*: What is Pr (model is true)?
- *Not*: Can we "reject" the model?
- Expanding the model
- Including more data

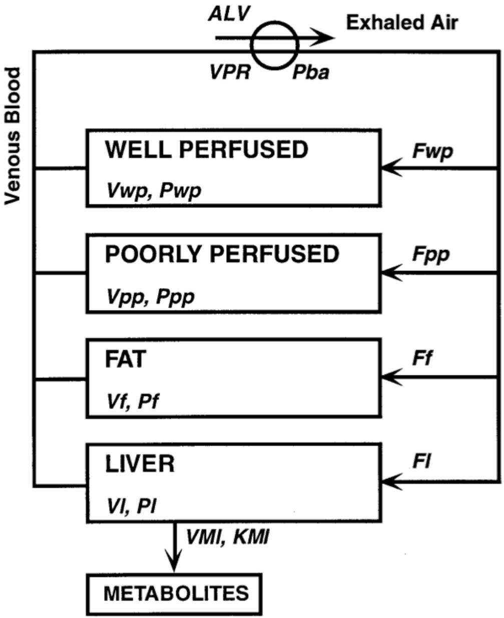
Example 2: Toxicology

- Central story: 4-compartment model of toxicokinetics of perchloroethylene
- Bayesian inference combines prior information and data
- Unresolved questions
- How the model all fits together

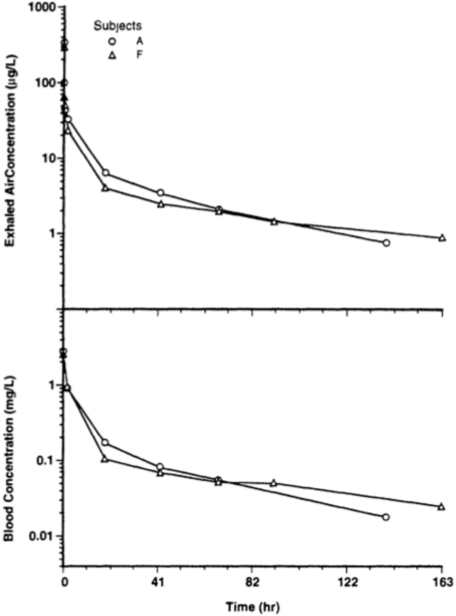
Toxicokinetics of perchloroethylene

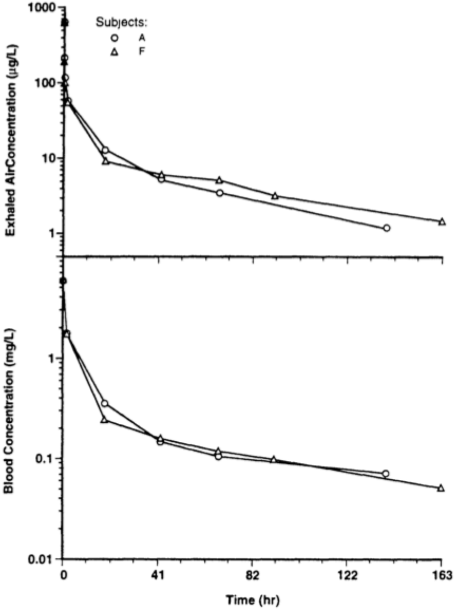
- Goal:
  - How much PERC is metabolized at low doses
  - Population distribution
- Experimental data: Expose 6 healthy volunteers to PERC for four hours, then measure concentrations in blood and air for 2 weeks
- 4-compartment model, metabolism in liver
- Our analysis:
  - Simple data-fitting did not work
  - Use Bayes to combine data and prior info within model

4-compartment model



Some data


1000
100
10
1
Subjects
A
F
Exhaled AirConcentration (µg/L)
1
0.1
0.01
Blood Concentration (mg/L)
0
41
82
122
163
Time (hr)


1000
100
10
1
Subjects:
A
F
Exhaled AirConcentration (µg/L)
1
0.1
0.01
Blood Concentration (mg/L)
0
41
82
122
163
Time (hr)

Simple statistical ideas did not work

- Fitting 4-compartment model directly to data
- Assisted model fit
- 1 or 2-compartment model
- Simulation from prior distribution

Simple statistical ideas that did not work: Fitting 4-compartment model directly to data

- Nonlinear least squares
- Fitting to each person separately:
    - Unstable: approx 30 data points, 15 param
    - “8 kg liver”
- Pooling data and estimating parameters for “the standard man”
    - Not useful for our goal of population inference

Simple statistical ideas that did not work: Assisted model fit

- Set some parameters to fixed values (from the pharmacology literature)
- Estimate the other parameters
- Results:
  - Couldn't fit the data well
  - Difficult to get fixed values for PERC-specific parameters such as equilibrium concentration ratios

Simple statistical ideas that did not work: 1 or 2-compartment model

- Simpler model can be estimated easily and robustly
- Does not fit the data well
  - Most of the PERC leaves in a few hours, but some stays in the body after a week or more
- Not realistic for low-dose extrapolation

Simple statistical ideas that did not work: Simulation from prior distribution

- Get prior information on parameters from pharmacology literature
- Try to fit data within these prior constraints
- Does not fit the data well
- Difficult to get good prior information for PERC-specific parameters such as equilibrium concentration ratios

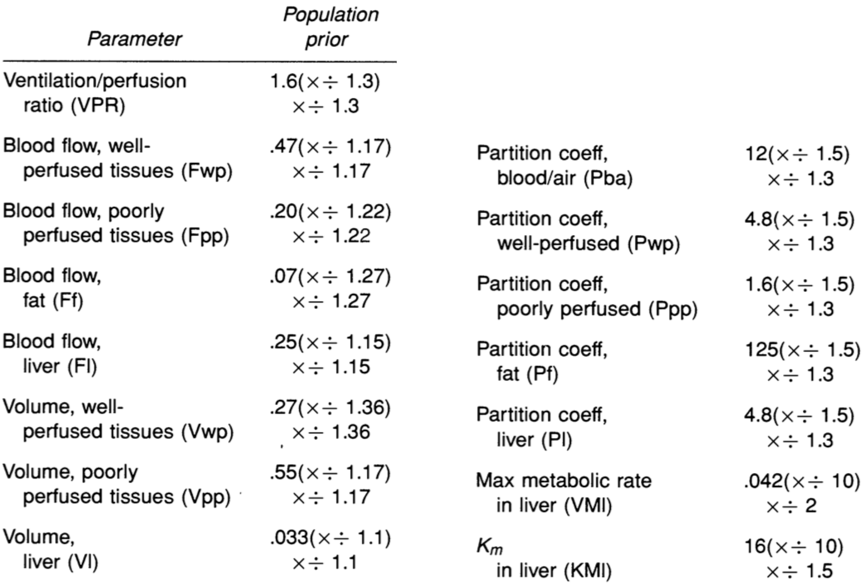
Bayesian inference

- 4-compartment model
- 15 parameters for each person
- Prior information
  - Strong for some parameters (e.g., volume of liver)
  - Weak for others (e.g., Michaelis-Menten coef)
  - Model includes uncertainty and variation
- Posterior simulation: random walk through parameter space
- Inference for parameters and predictions
- Model checking

Hierarchical prior distributions

- Prior distribution for a rate parameter in the metabolism, θ_j for person j
  - $\log \theta_j \sim N(\mu, \tau^2)$
  - $\mu \sim N(\log 16, (\log 10)^2)$
  - $\tau \approx \log 2$
- Large uncertainty, small variation
- Can learn about μ using data from several people
- Can't do this without a hierarchical model
- Transformations and prior correlations (why transformations are particularly important for Bayesians)

Hierarchical prior distributions



Parameter	*Population prior*
Ventilation/perfusion ratio (VPR)	1.6(×÷ 1.3) ×÷ 1.3
Blood flow, well-perfused tissues (Fwp)	.47(×÷ 1.17) ×÷ 1.17
Blood flow, poorly perfused tissues (Fpp)	.20(×÷ 1.22) ×÷ 1.22
Blood flow, fat (Ff)	.07(×÷ 1.27) ×÷ 1.27
Blood flow, liver (Fl)	.25(×÷ 1.15) ×÷ 1.15
Volume, well-perfused tissues (Vwp)	.27(×÷ 1.36) ×÷ 1.36
Volume, poorly perfused tissues (Vpp)	.55(×÷ 1.17) ×÷ 1.17
Volume, liver (Vl)	.033(×÷ 1.1) ×÷ 1.1
Partition coeff, blood/air (Pba)	12(×÷ 1.5) ×÷ 1.3
Partition coeff, well-perfused (Pwp)	4.8(×÷ 1.5) ×÷ 1.3
Partition coeff, poorly perfused (Ppp)	1.6(×÷ 1.5) ×÷ 1.3
Partition coeff, fat (Pf)	125(×÷ 1.5) ×÷ 1.3
Partition coeff, liver (Pl)	4.8(×÷ 1.5) ×÷ 1.3
Max metabolic rate in liver (VMl)	.042(×÷ 10) ×÷ 2
K_m in liver (KMl)	16(×÷ 10) ×÷ 1.5

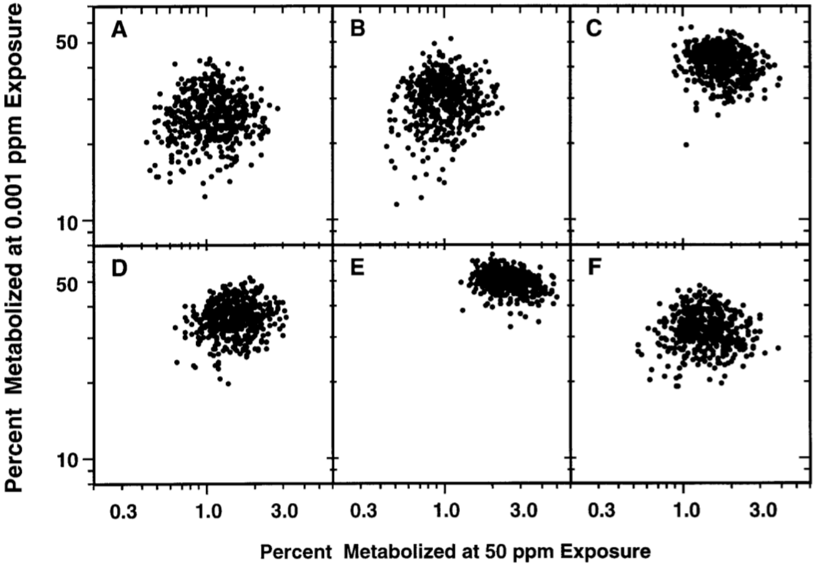
What we did

- Set up a hierarchical prior distribution with uncertainty and population variation for a 4-compartment model
- Fit the model to data (much computation)
- Checked inferences about parameters to see that they made sense
- Re-ran model under hypothetical low-dose, high-dose exposures

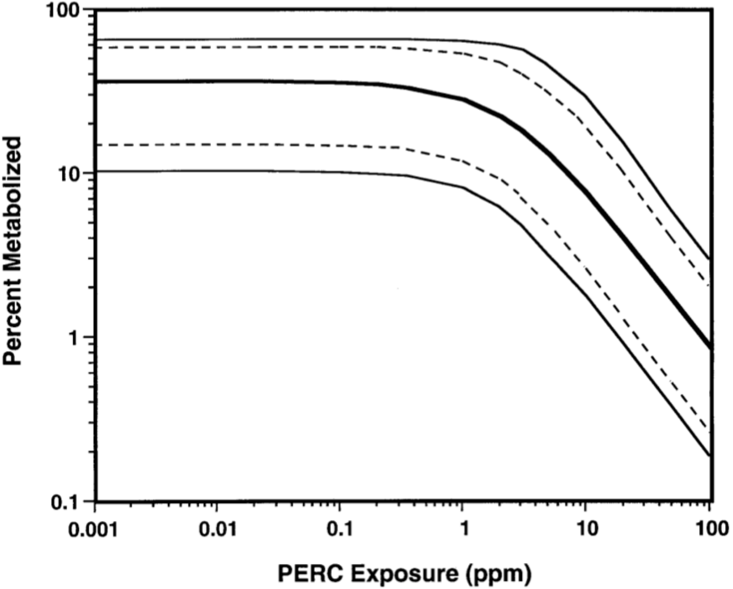
Fitting and using the model

- Use Gibbs sampler and Metropolis algorithm to take a random walk through parameter space
- Computationally intensive
  - Each step requires evaluation of the numerical differential equation solver
- Check inferences: Do they make sense?
- Re-run the model several times to simulate what would happen under different conditions

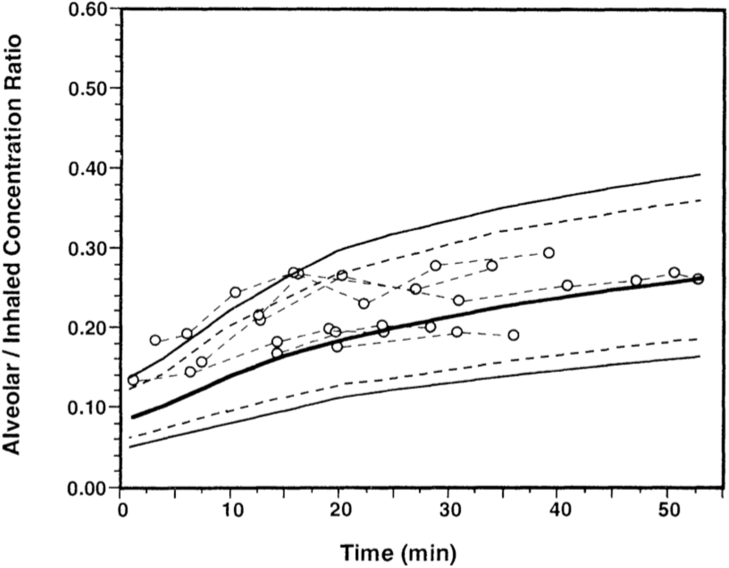
Inference for 6 individuals



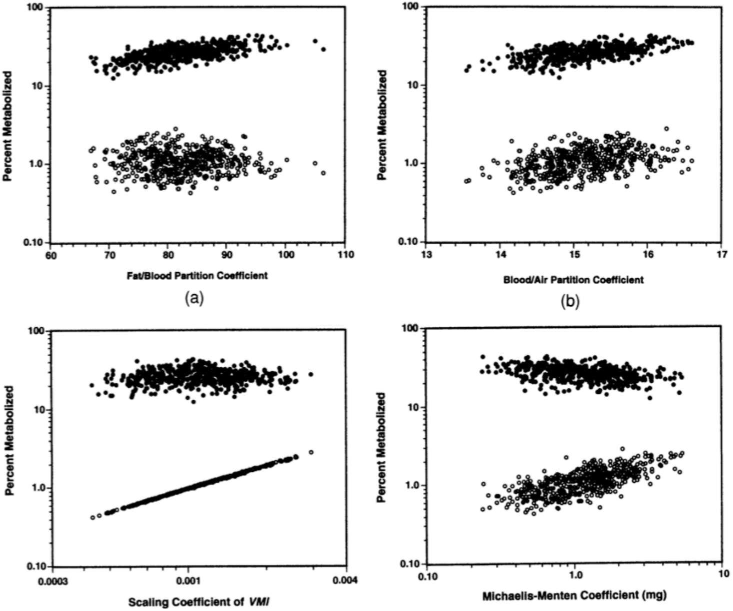
Inference for the population



Prediction of data from a new study



Sensitivity to priors



(a)

(b)

Putting it all together

(a) Physiological pharmacokinetic model
(b) Hierarchical population model
(c) Prior information
(d) Experimental data
(e) Bayesian inference
(f) Computation
(g) Model checking

- We need all of these!

(a) Physiological pharmacokinetic model

- Without a physiological model, there is no good way to get prior information on the parameters
- We need physiological parameters (not just curve-fitting of the data) to efficiently combine information across different people

(b) Hierarchical population model

- Without a population model, there generally are not enough data to estimate the parameters separately for each individual
- And there is too much variation among bodies (even among healthy young male volunteers) to pool all the data together and estimate common parameters

(c) Prior information
(d) Experimental data

- We need prior information. Otherwise, our estimates don't make sense (the 8 kg liver)
- We need experimental data to learn about perchloroethylene in particular

(e) Bayesian inference

- Using Bayesian inference, we can find parameters that are consistent with both prior information and data, if such agreement is possible
- Automatically includes uncertainty and variability, so inferences can be plugged in directly to risk assessment and decision analysis

(f) Computation

- Our models are big. Least squares, maximum likelihood, etc., are not enough
- Old-fashioned differential-equation solver is still sitting inside the model
- Our computers are never fast enough. We want more, more, more!

(g) Model checking

- Check inferences about parameters
  - Do they make sense?
  - Are they consistent with prior distributions
- Check fit to data
- Check predictions on new data

Using Bayesian ideas to improve existing analyses

- Regularization (for example, avoiding estimates on the boundary of parameter space)
- Accounting for uncertainty (especially for decisions)
- Checking model fit
- Using models to combine different sources of information (partial pooling)
- Better dialogue with subject-matter experts (more windows into the model and data)

Summary of example

- Population pharmacokinetic models have many moving parts
- Complexity in one place can make it *easier*, not harder, to add information in other places

“Bayesian Data Analysis”

- “Bayesian inference” is too narrow; “Bayesian statistics” is too broad
- “Bayes” is a good brand name; “Statistics using conditional probability” is confusing
- Everyone uses Bayesian inference when it is appropriate.
  A Bayesian is a statistician who uses Bayesian inference even when it is inappropriate.
  I am a Bayesian.

What is Bayes?

- Bayes is data + regularization
- Bayes is data + prior information
- Bayes is logical probabilistic reasoning
- Bayes is different things at different times

The three steps of Bayesian data analysis

- Three steps:
    1. Setting up a probability model
    2. Inference
    3. Model checking
- Then go back and improve the model

Example 3: Spell checking

- The typed word “Radom” is actually Random ($\theta = 1$), Radon ($\theta = 2$), or Radom ($\theta = 3$)
- Prior distribution:

θ	$p(\theta)$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

- Likelihood:

θ	$p(\text{“radom”}\,\vert\theta)$
random	0.00193
radon	0.000143
radom	0.975

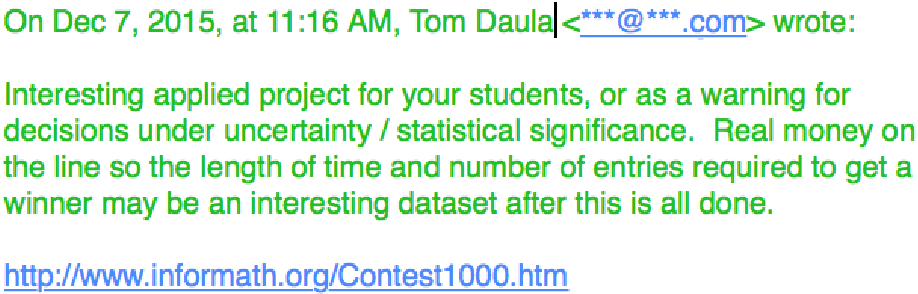
Spell checking

- Prior, likelihood, posterior distributions:

θ	$p(\theta)$	$p(y\|\theta)$	$p(\theta)p(y\|\theta)$	$p(\theta\|\text{y})$
random	7.60×10^{-5}	0.00193	1.47×10^{-7}	0.325
radon	6.05×10^{-6}	0.000143	8.65×10^{-10}	0.002
radom	3.12×10^{-7}	0.975	3.04×10^{-7}	0.673

- Decision making
- Model checking
- Model improvement

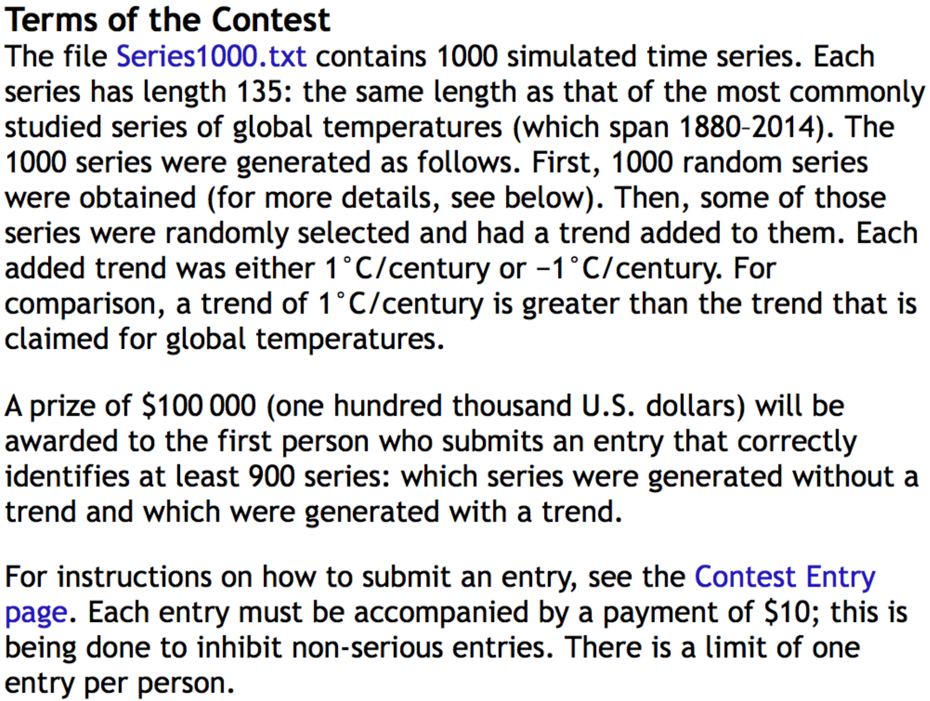
Example 4: “Global climate challenge”



On Dec 7, 2015, at 11:16 AM, Tom Daula <***@***.com> wrote:

Interesting applied project for your students, or as a warning for decisions under uncertainty / statistical significance. Real money on the line so the length of time and number of entries required to get a winner may be an interesting dataset after this is all done.

http://www.informath.org/Contest1000.htm



Terms of the Contest

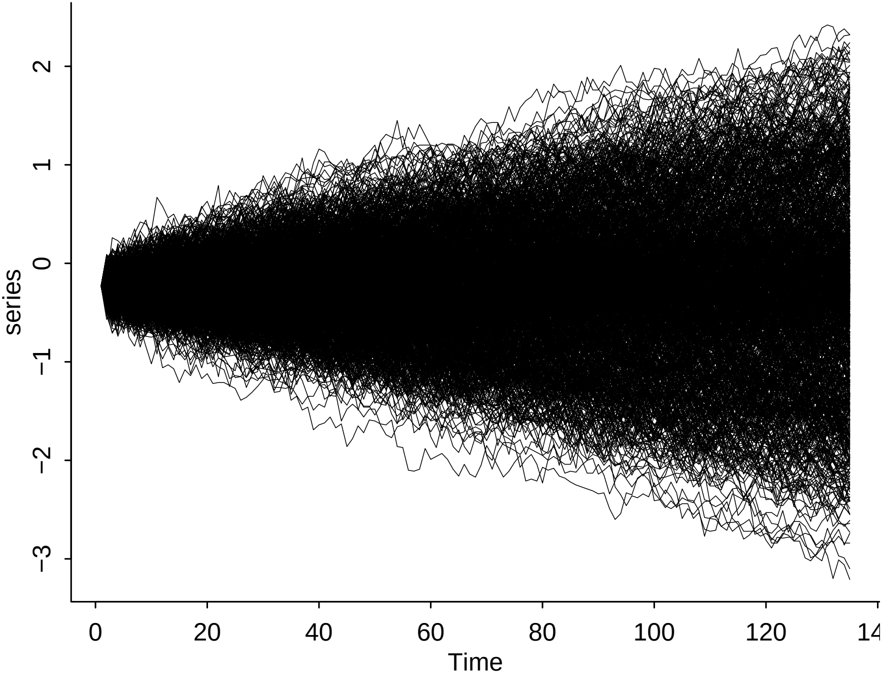
The file Series1000.txt contains 1000 simulated time series. Each series has length 135: the same length as that of the most commonly studied series of global temperatures (which span 1880–2014). The 1000 series were generated as follows. First, 1000 random series were obtained (for more details, see below). Then, some of those series were randomly selected and had a trend added to them. Each added trend was either 1 °C/century or −1 °C/century. For comparison, a trend of 1 °C/century is greater than the trend that is claimed for global temperatures.

A prize of $100 000 (one hundred thousand U.S. dollars) will be awarded to the first person who submits an entry that correctly identifies at least 900 series: which series were generated without a trend and which were generated with a trend.

For instructions on how to submit an entry, see the Contest Entry page. Each entry must be accompanied by a payment of $10; this is being done to inhibit non-serious entries. There is a limit of one entry per person.

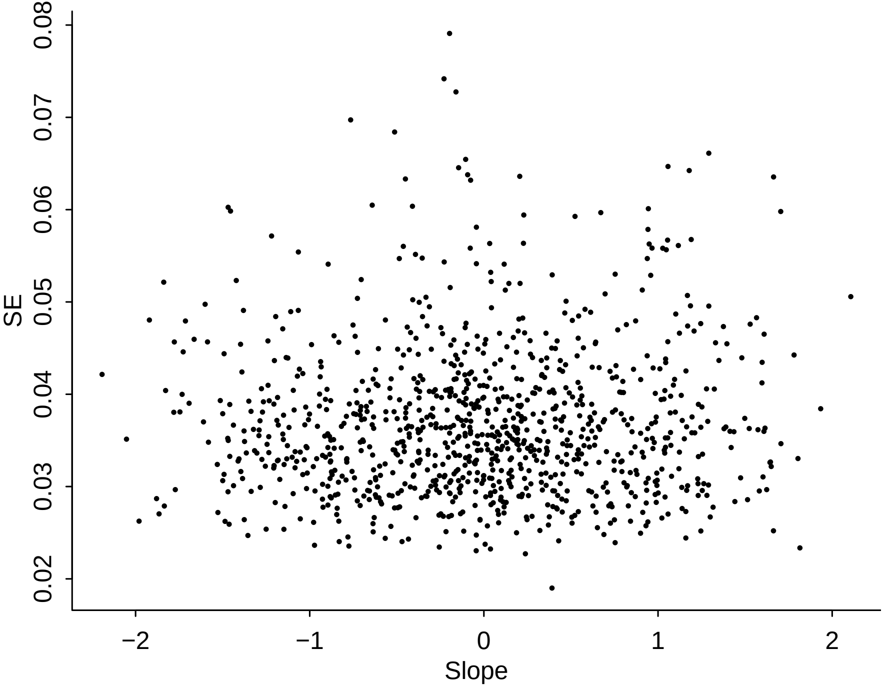
Download and graph the data

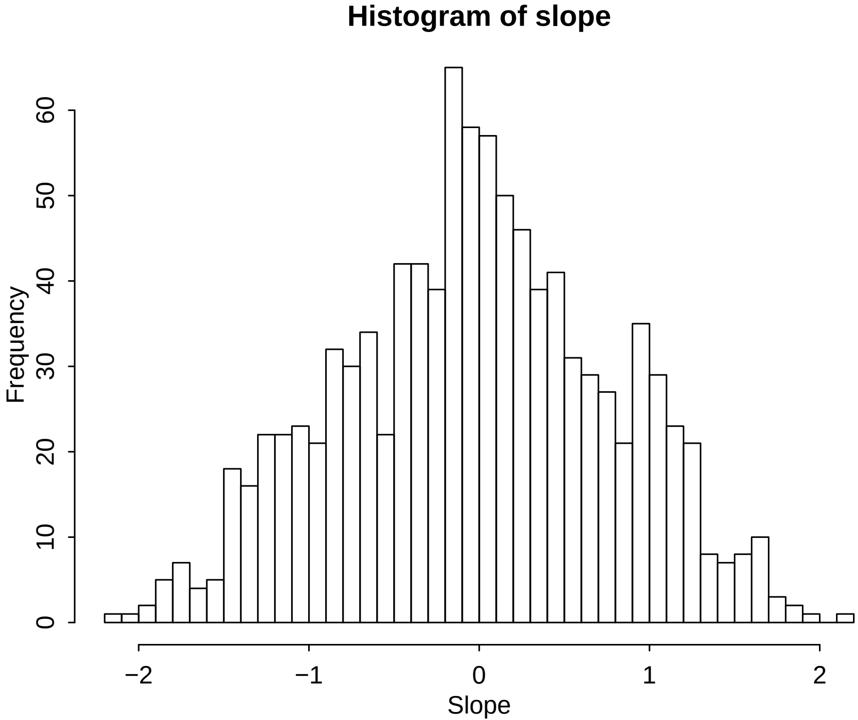
```
series <- matrix(scan("Series1000.txt"), nrow=1000, ncol=135,
  byrow=TRUE)
T <- 135
N <- 1000

pdf("series_1.pdf", height=5, width=6)
par(mar=c(3,3,2,0), tck=-.01, mgp=c(1.5,.5,0))
plot(c(1,T), range(series), bty="l", type="n",
  xlab="Time", ylab="series")
for (n in 1:N){
  lines(1:T, series[n,], lwd=.5)
}
dev.off()
```


series
2
1
0
−1
−2
−3
Time
0
20
40
60
80
100
120
14

Fit a regression to each line and plot the estimated slopes

```
library("arm")
slope <- rep(NA, N)
se <- rep(NA, N)
for (n in 1:N){
  data <- series[n,]
  time <- 1:T
  fit <- lm(data ~ time)
  slope[n] <- 100*coef(fit)[2]
  se[n] <- 100*se.coef(fit)[2]
}
pdf("series_2.pdf", height=5, width=6)
par(mar=c(3,3,2,0), tck=-.01, mgp=c(1.5,.5,0))
plot(slope, se, bty="l", xlab="Slope", ylab="SE",pch=20,cex=.5)
dev.off()
```


SE
0.02
0.03
0.04
0.05
0.06
0.07
0.08
−2
−1
0
1
2
Slope


Histogram of slope
Frequency
0
10
20
30
40
50
60
−2
−1
0
1
2
Slope

Program a mixture model in Stan

```
data {
  int K;
  int N;
  real y[N];
  real mu[K];
}
parameters {
  simplex[K] theta;
  real sigma;
}
model {
  real ps[K];
  sigma ~ cauchy(0,2.5);
  mu ~ normal(0,10);
  for (n in 1:N) {
    for (k in 1:K) {
      ps[k] <- log(theta[k]) + normal_log(y[n], mu[k], sigma);
    }
    increment_log_prob(log_sum_exp(ps));
  }
```

Run the model in R

```
y <- slope
K <- 3
mu <- c(0,-1,1)
mix <- stan("mixture.stan")
print(mix)

>
Inference for Stan model: mixture.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

         mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
theta[1] 0.54       0 0.02 0.49 0.52 0.54 0.55  0.58  2449    1
theta[2] 0.24       0 0.02 0.21 0.23 0.24 0.25  0.27  2537    1
theta[3] 0.22       0 0.02 0.19 0.21 0.22 0.23  0.26  2444    1
sigma    0.40       0 0.02 0.37 0.39 0.40 0.42  0.45  2078    1
```

For each series, compute probability of it being in each component

```
generated quantities {
  matrix[N,K] p;
  for (n in 1:N){
    vector[K] p_raw;
    for (k in 1:K){
      p_raw[k] <- theta[k]*exp(normal_log(y[n], mu[k], sigma));
    }
    for (k in 1:K){
      p[n,k] <- p_raw[k]/sum(p_raw);
    }
  }
}
```

Results

```
      [,1] [,2] [,3]
 [1,] 0.09 0.00 0.91
 [2,] 0.41 0.59 0.00
 [3,] 0.93 0.01 0.06
 [4,] 0.83 0.17 0.00
 [5,] 0.82 0.17 0.00
 [6,] 0.95 0.01 0.05
 [7,] 0.74 0.00 0.26
 [8,] 0.86 0.14 0.00
 [9,] 0.11 0.00 0.89
[10,] 0.87 0.00 0.13
[11,] 0.94 0.01 0.06
[12,] 0.29 0.71 0.00
[13,] 0.09 0.91 0.00
[14,] 0.67 0.33 0.00
[15,] 0.93 0.01 0.06
[16,] 0.95 0.01 0.04
[17,] 0.16 0.84 0.00
[18,] 0.95 0.04 0.01
[19,] 0.77 0.23 0.00
```

Summaries

- Best guess for each series:

```
  1   2   3
559 232 209
```

- Expected # correct and sd:

```
854.1  10.3
```

- Probability of getting at least 900 correct:

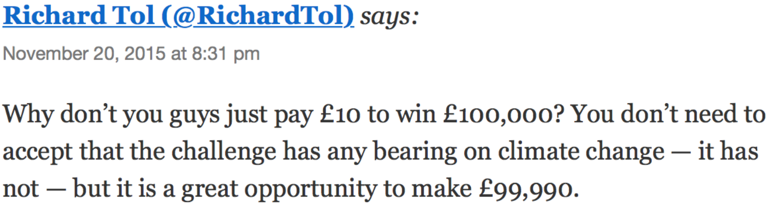
```
> pnorm(expected_correct, 899.5, sd_correct)
[1] 5.421277e-06
```

- Ummmmm ...

```
> 1/pnorm(expected_correct, 899.5, sd_correct)
[1] 184458.4
```

Should I play the $100,000 challenge?

- From the discussion thread:



Richard Tol (@RichardTol) *says:*

November 20, 2015 at 8:31 pm

Why don't you guys just pay £10 to win £100,000? You don't need to accept that the challenge has any bearing on climate change — it has not — but it is a great opportunity to make £99,990.

- Expected return on $10 bet:

$$(5.4 \cdot 10^{-6}) \cdot 10^{5} = \$0.54$$

- What would *you* do?

Bayesian data analysis in context

Different approaches to statistics:

- Traditional likelihood
- Pure nonparametric, robust
- Full Bayes modeling
  - “A chicken is an egg’s way of making another egg”