Bayesian Statistics without Frequentist Language

Richard McElreath
Max Planck Institute for Evolutionary Anthropology
Leipzig

Outside view

Likelihood function

From Wikipedia, the free encyclopedia

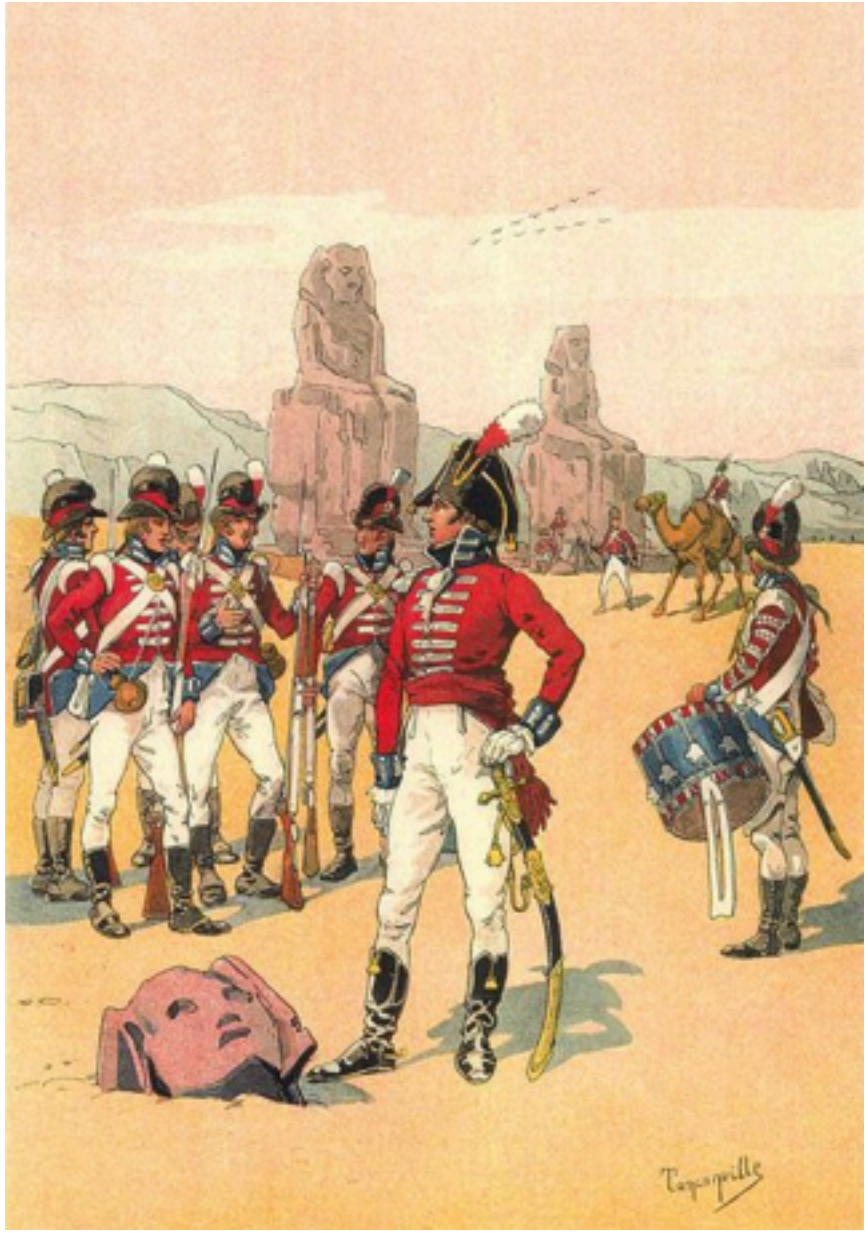
> *For statistical inference using likelihood functions, see Bayesian statistics, maximum-likelihood estimation, and likelihood-ratio testing.*

In statistics, a **likelihood function** (often simply the **likelihood**) is a function of the parameters of a statistical model given data. Likelihood functions play a key role in statistical inference, especially methods of estimating a parameter from a set of statistics. In informal contexts, "likelihood" is often used as a synonym for "probability." In statistics, a distinction is made depending on the roles of outcomes vs. parameters. *Probability* is used before data are available to describe possible future outcomes given a fixed value for the parameter (or parameter vector). *Likelihood* is used after data are available to describe a function of a parameter (or parameter vector) for a given *outcome*.

R.A. Fisher (1890–1962)

Outside view

- Data have distributions
- Parameters do not
- Distinguish *parameters* and *statistics*
- Likelihood not a probability distribution
- Imaginary *population*
- Bayes is sampling theory + priors
- Priors are uniquely subjective



Lineage of complaints

What most statisticians have is a parody of the Bayesian argument, a simplistic view that just adds a woolly prior to the sampling-theory paraphernalia. They look at the parody, see how absurd it is, and thus dismiss the coherent approach as well. Efron has studied the Bayesian argument more than have most statisticians, but it is still only a parody that is presented in this article. Many of the arguments he produces are distortions of the thing he is attacking.

Dennis Lindley (1923–2013)

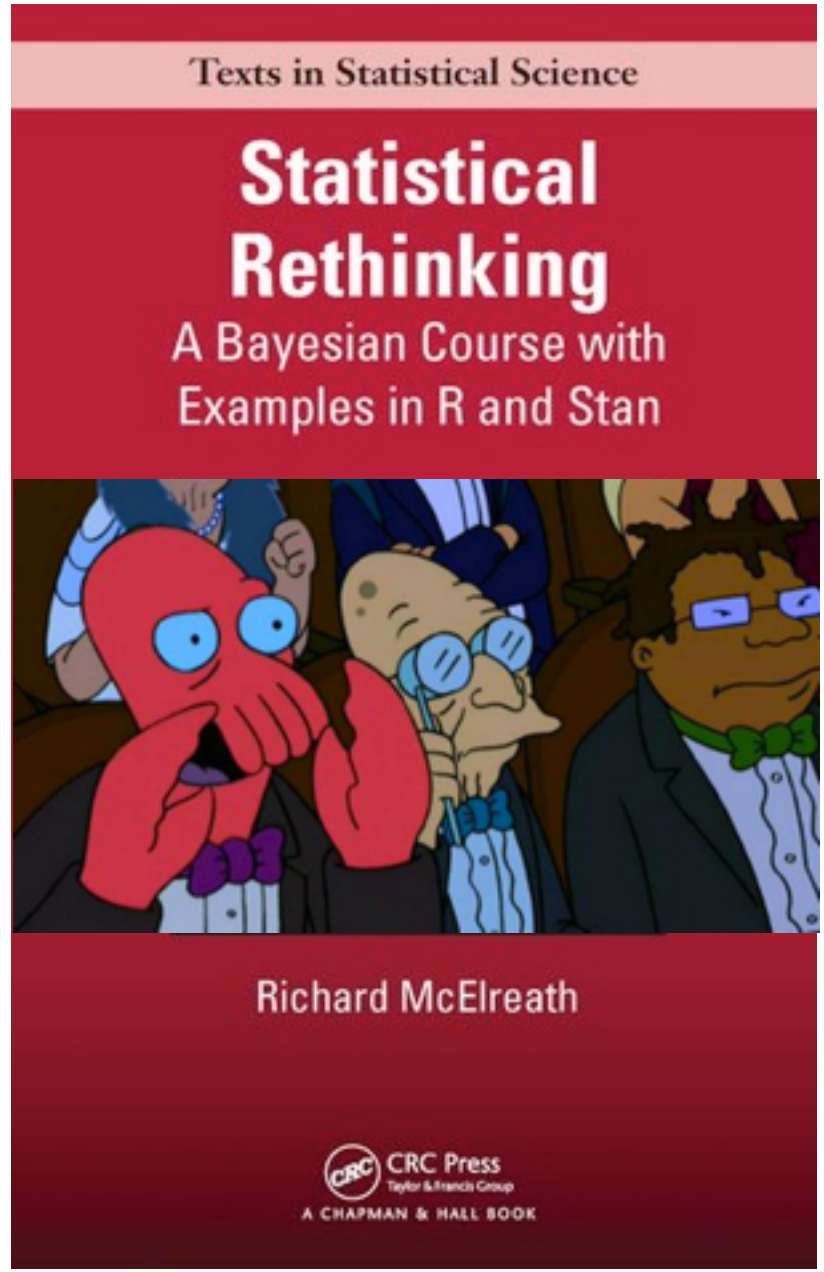
The American Statistician, February 1986, Vol. 40, No. 1

Conceptual friction

- Common barriers:
  - Thinking data must look like likelihood function
  - Degrees of freedom
  - "Sampling" as source of all uncertainty
  - Defining random effects via sampling design
  - Neglect of data uncertainty
  - add your own

My Book is Neo-Colonial

- I feel bad about choices made
- Uses outsider perspective
  - "Likelihood"
  - "parameter"
  - "estimate"
- Like explaining Indian politics using British political parties
- Perpetuates confusion
- Historical necessity?



Another path

- Claim: Bayes easier and more powerful when understood from the inside
- Problem: Many insider views

46656 Varieties of Bayesians (#765)

Some attacks and defenses of the Bayesian position assume that it is unique so it should be helpful to point out that there are at least 46656 different interpretations. This is shown by the following classification based on eleven facets. The count would be larger if I had not artificially made some of the facets discrete and my heading would have been "On the Infinite Variety of Bayesians."

I.J. Good 1971

Insider perspective

- Bayesian approach: A joint generative model of all variables

- Key ideas:
  - *Unity among variables*: No deep distinction between data and parameters
  - *Unity among distributions*: No deep distinction between likelihoods and priors

Likelihood or Prior?

$$\beta \sim \text{Normal}(\flat, \lightning)$$

Likelihood or Prior?

$$\beta \sim \text{Normal}(\beta, \lightning)$$

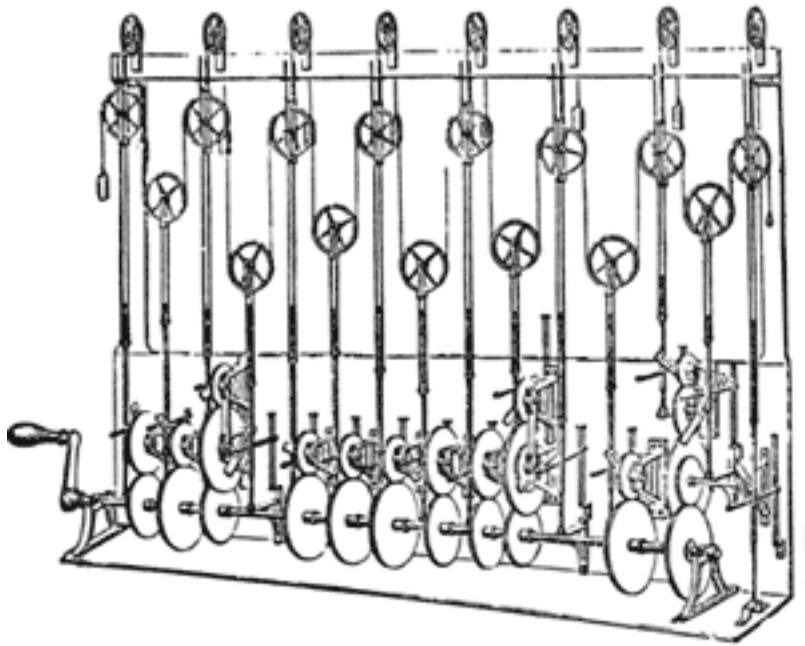
If β is observed, likelihood.

If β is unobserved, prior.

Corner cases

- In conventional GLMs, no problem distinguishing data from parameters.

- But what about:
  - GLMMs
  - Missing data
  - Measurement error
  - *Many strange machines*



notes

rate of singing when cat present

rate of singing when cat absent

cat

Observed variables

`notes`

`cat`

Unobserved variables

rate of singing when cat present

rate of singing when cat absent

Joint model

Prob(notes, cat, *rate*|cat, *rate*|no-cat)

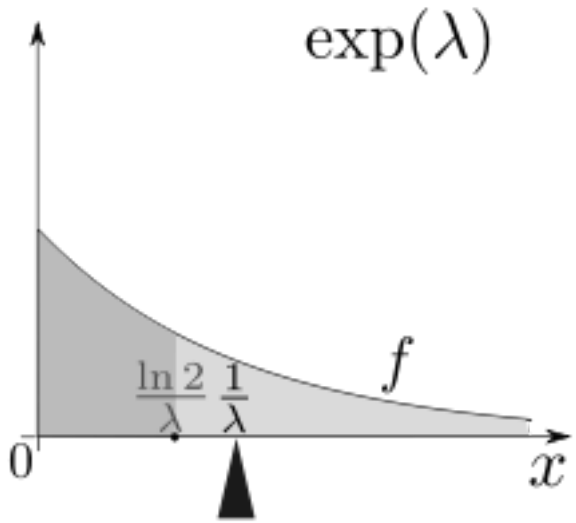
Joint model

Prob(notes, cat, *rate*|cat, *rate*|no-cat)

$$
\begin{aligned}
\text{notes}_t &\sim \text{Poisson}(\lambda_t) \\
\lambda_t &= (1 - \text{cat}_t)\alpha + \text{cat}_t\beta \\
\alpha &\sim \text{Exponential}(1/10) \\
\beta &\sim \text{Exponential}(1/10)
\end{aligned}
$$

How is prior formed?

- What pre-data information do we have about unobserved variables?
  - Rates are non-zero positive real values. Model expected value ==maxent==> Exponential
  - This most conservative distribution consistent w info



How is prior formed?

- What pre-data information do we have about unobserved variables?
  - Rates are non-zero positive real values. Model expected value ==maxent==> Exponential
  - This most conservative distribution consistent w info
- Like priors, likelihoods are pre-data distributions.
  - Use pre-data information (meta-data) to build them.
  - Notes are zero or positive integers. Model expected value ==maxent==> Poisson
  - Again, most conservative distribution consistent w info

$$\text{notes}_t \sim \text{Poisson}(\lambda_t)$$
$$\lambda_t = (1 - \text{cat}_t)\alpha + \text{cat}_t\beta$$
$$\alpha \sim \text{Exponential}(1/10)$$
$$\beta \sim \text{Exponential}(1/10)$$

Stan code

```
data{
    int<lower=1> N;
    int notes[N];
    int cat[N];
}
parameters{
    real<lower=0> alpha;
    real<lower=0> beta;
}
model{
    vector[N] lambda;
    beta ~ exponential( 0.1 );
    alpha ~ exponential( 0.1 );
    for ( i in 1:N ) {
        lambda[i] = (1 - cat[i]) * alpha + cat[i] * beta;
    }
    notes ~ poisson( lambda );
}
```

map2stan code

```
notes ~ poisson(lambda),
lambda <- (1-cat)*alpha + cat*beta,
alpha ~ exponential(0.1),
beta ~ exponential(0.1)
```

https://gist.github.com/rmcelreath

GLMM birds

- Multiple birds, each with own rates:

$$
\begin{aligned}
\text{notes}_{it} &\sim \text{Poisson}(\lambda_{it}) \\
\lambda_{it} &= (1 - \text{cat}_{it})\alpha_i + \text{cat}_{it}\beta_i \\
\alpha_i &\sim \text{Exponential}(1/\bar{\alpha}) \\
\beta_i &\sim \text{Exponential}(1/\bar{\beta}) \\
\bar{\alpha} &\sim \text{Exponential}(1/10) \\
\bar{\beta} &\sim \text{Exponential}(1/10)
\end{aligned}
$$

The Annals of Statistics
2005, Vol. 33, No. 1, 1–53
DOI 10.1214/009053604000001048

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$. Kreft and de Leeuw [(1998), page 12] thus distinguish between fixed and random coefficients.

The Annals of Statistics
2005, Vol. 33, No. 1, 1–53
DOI 10.1214/009053604000001048

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$. Kreft and de Leeuw [(1998), page 12] thus distinguish between fixed and random coefficients.
2. Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population. Searle, Casella and McCulloch [(1992), Section 1.4] explore this distinction in depth.

The Annals of Statistics
2005, Vol. 33, No. 1, 1–53
DOI 10.1214/009053604000001048

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$. Kreft and de Leeuw [(1998), page 12] thus distinguish between fixed and random coefficients.
2. Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population. Searle, Casella and McCulloch [(1992), Section 1.4] explore this distinction in depth.
3. "When a sample exhausts the population, the corresponding variable is *fixed*; when the sample is a small (i.e., negligible) part of the population the corresponding variable is *random*" [Green and Tukey (1960)].

The Annals of Statistics
2005, Vol. 33, No. 1, 1–53
DOI 10.1214/009053604000001048

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$. Kreft and de Leeuw [(1998), page 12] thus distinguish between fixed and random coefficients.
2. Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population. Searle, Casella and McCulloch [(1992), Section 1.4] explore this distinction in depth.
3. "When a sample exhausts the population, the corresponding variable is *fixed*; when the sample is a small (i.e., negligible) part of the population the corresponding variable is *random*" [Green and Tukey (1960)].
4. "If an effect is assumed to be a realized value of a random variable, it is called a random effect" [LaMotte (1983)].

The Annals of Statistics
2005, Vol. 33, No. 1, 1–53
DOI 10.1214/009053604000001048

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$. Kreft and de Leeuw [(1998), page 12] thus distinguish between fixed and random coefficients.
2. Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population. Searle, Casella and McCulloch [(1992), Section 1.4] explore this distinction in depth.
3. "When a sample exhausts the population, the corresponding variable is *fixed*; when the sample is a small (i.e., negligible) part of the population the corresponding variable is *random*" [Green and Tukey (1960)].
4. "If an effect is assumed to be a realized value of a random variable, it is called a random effect" [LaMotte (1983)].
5. Fixed effects are estimated using least squares (or, more generally, maximum likelihood) and random effects are estimated with shrinkage ["linear unbiased prediction" in the terminology of Robinson (1991)]. This definition is standard in the multilevel modeling literature [see, e.g., Snijders and Bosker (1999), Section 4.2] and in econometrics.

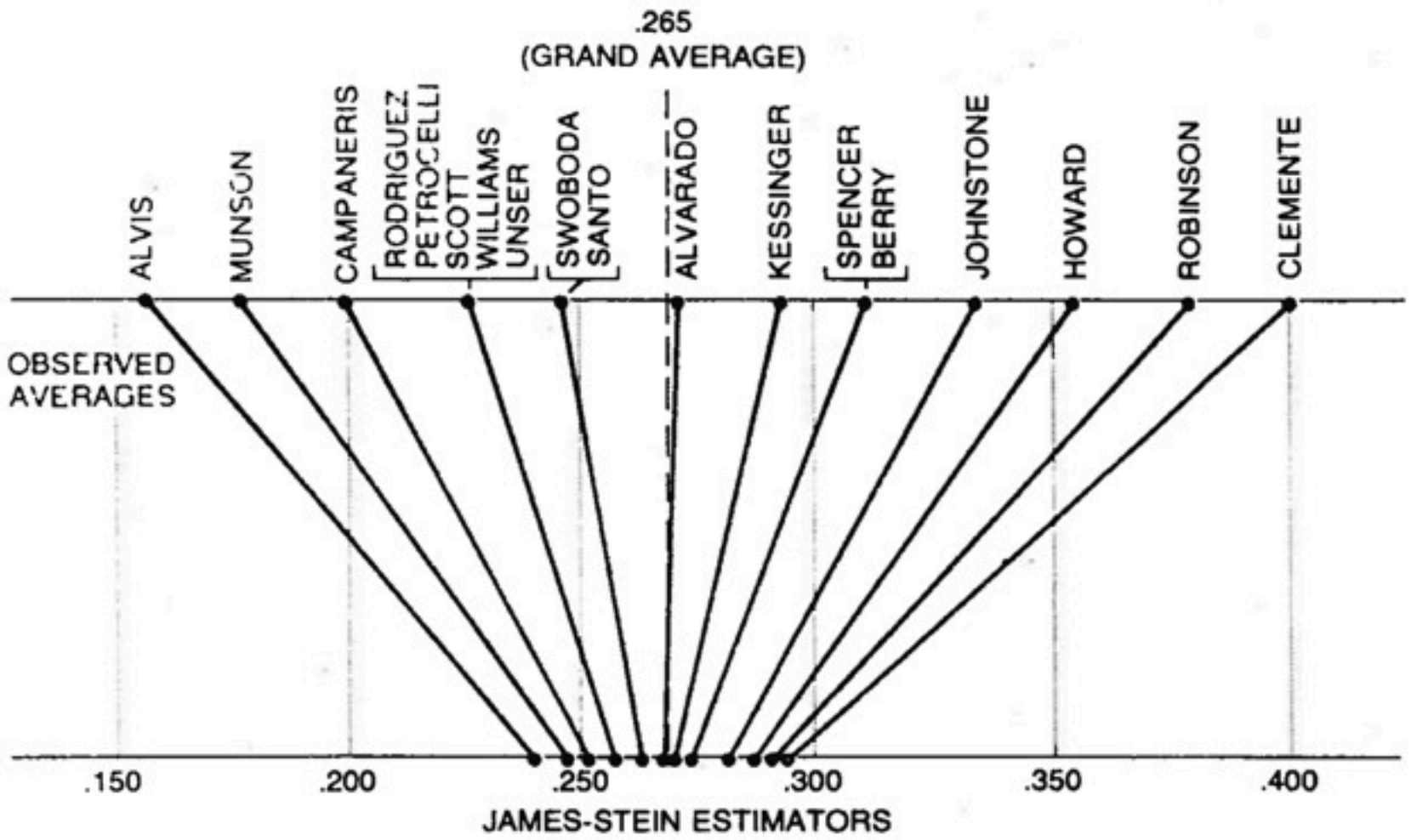
   In the Bayesian framework, this definition implies that fixed effects $\beta_j^{(m)}$ are estimated conditional on $\sigma_m = \infty$ and random effects $\beta_j^{(m)}$ are estimated conditional on σ_m from the posterior distribution.

GLMM birds

- *Shrinkage* happens everywhere

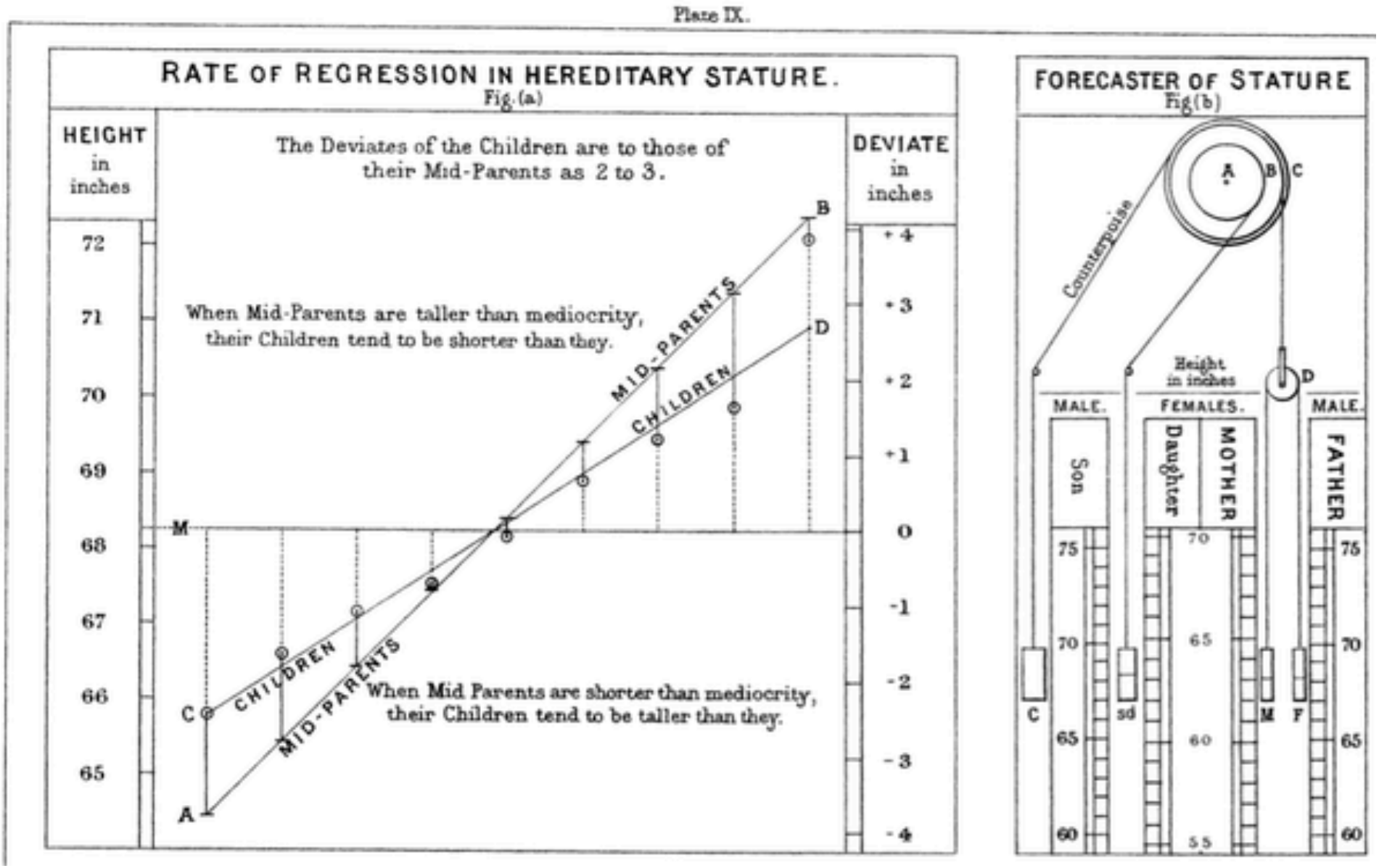
$$
\begin{aligned}
\text{notes}_{it} &\sim \text{Poisson}(\lambda_{it}) \\
\lambda_{it} &= (1 - \text{cat}_{it})\alpha_i + \text{cat}_{it}\beta_i \\
\alpha_i &\sim \text{Exponential}(1/\bar{\alpha}) \\
\beta_i &\sim \text{Exponential}(1/\bar{\beta}) \\
\bar{\alpha} &\sim \text{Exponential}(1/10) \\
\bar{\beta} &\sim \text{Exponential}(1/10)
\end{aligned}
$$

Efron's example of "shrinkage estimator"



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Galton's "regression to mean"



$$\text{notes}_{it} \sim \text{Poisson}(\lambda_{it})$$
$$\lambda_{it} = (1 - \text{cat}_{it})\alpha_i + \text{cat}_{it}\beta_i$$
$$\alpha_i \sim \text{Exponential}(1/\bar{\alpha})$$
$$\beta_i \sim \text{Exponential}(1/\bar{\beta})$$
$$\bar{\alpha} \sim \text{Exponential}(1/10)$$
$$\bar{\beta} \sim \text{Exponential}(1/10)$$

Stan code

```
data{
    int<lower=1> N;
    int<lower=1> N_id;
    int notes[N];
    int cat[N];
    int id[N];
}
parameters{
    vector<lower=0>[N_id] alpha;
    vector<lower=0>[N_id] beta;
    real<lower=0> alpha_bar;
    real<lower=0> beta_bar;
}
model{
    vector[N] lambda;
    beta_bar ~ exponential( 0.1 );
    alpha_bar ~ exponential( 0.1 );
    beta ~ exponential( 1.0/beta_bar );
    alpha ~ exponential( 1.0/alpha_bar );
    for ( i in 1:N ) {
        lambda[i] = (1 - cat[i]) * alpha[id[i]]
                    + cat[i] * beta[id[i]];
    }
    notes ~ poisson( lambda );
}
```

map2stan code

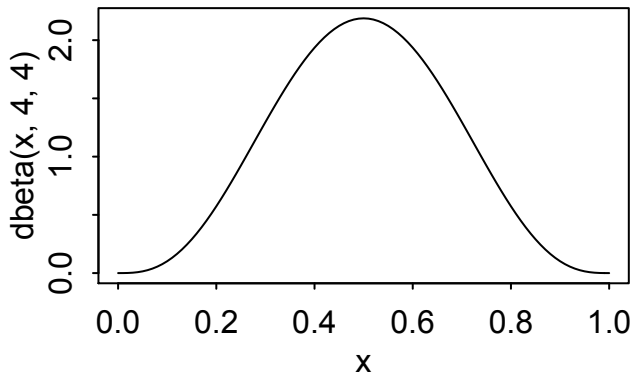
```
notes ~ poisson(lambda),
lambda <- (1-cat)*alpha[id] + cat*beta[id],
alpha[id] ~ exponential(1.0/alpha_bar),
beta[id] ~ exponential(1.0/beta_bar),
alpha_bar ~ exponential(0.1),
beta_bar ~ exponential(0.1)
```

https://gist.github.com/rmcelreath

Bad data, good cats

- Jointly model cat behavior:

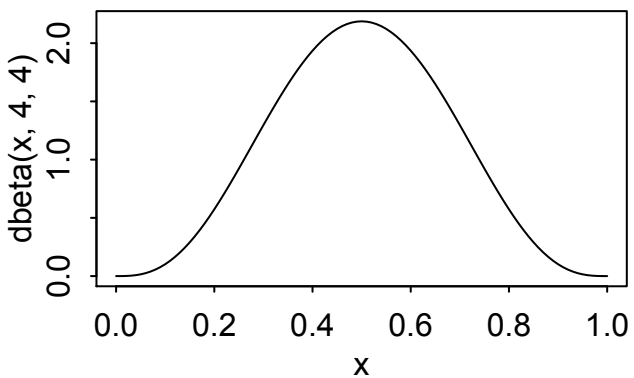
$$
\begin{aligned}
\text{notes}_t &\sim \text{Poisson}(\lambda_t) \\
\lambda_t &= (1 - \text{cat}_t)\alpha + \text{cat}_t\beta \\
\text{cat}_t &\sim \text{Bernoulli}(\kappa) \\
\kappa &\sim \text{Beta}(4, 4) \\
\alpha &\sim \text{Exponential}(1/10) \\
\beta &\sim \text{Exponential}(1/10)
\end{aligned}
$$



Bad data, good cats

- Useful when some data go missing: some cat_t observations unavailable—cats stepped on the keyboard.
- Same distribution does double duty:

$$
\begin{aligned}
\text{notes}_t &\sim \text{Poisson}(\lambda_t) \\
\lambda_t &= (1 - \text{cat}_t)\alpha + \text{cat}_t\beta \\
\text{cat}_t &\sim \text{Bernoulli}(\kappa) \\
\kappa &\sim \text{Beta}(4, 4) \\
\alpha &\sim \text{Exponential}(1/10) \\
\beta &\sim \text{Exponential}(1/10)
\end{aligned}
$$



$$\text{notes}_t \sim \text{Poisson}(\lambda_t)$$
$$\lambda_t = (1 - \text{cat}_t)\alpha + \text{cat}_t\beta$$
$$\text{cat}_t \sim \text{Bernoulli}(\kappa)$$
$$\kappa \sim \text{Beta}(4, 4)$$
$$\alpha \sim \text{Exponential}(1/10)$$
$$\beta \sim \text{Exponential}(1/10)$$

Stan code

```
parameters{
    real<lower=0,upper=1> kappa;
    real<lower=0> beta;
    real<lower=0> alpha;
}
model{
    beta ~ exponential( 0.1 );
    alpha ~ exponential( 0.1 );
    kappa ~ beta( 4 , 4 );
    for ( i in 1:N ) {
        if ( cat[i]==-1 ) { // cat missing
            target += log_mix( kappa ,
                    poisson_lpmf( notes[i] | beta ),
                    poisson_lpmf( notes[i] | alpha )
                );
        } else { // cat not missing
            cat[i] ~ bernoulli(kappa);
            notes[i] ~ poisson( (1-cat[i])*alpha +
cat[i]*beta );
        }
    }//i
}
```

map2stan code

```
notes ~ poisson(lambda),
lambda <- (1-cat)*alpha + cat*beta,
cat ~ bernoulli(kappa),
kappa ~ beta(4,4),
alpha ~ exponential(0.1),
beta ~ exponential(0.1)
```

https://gist.github.com/rmcelreath

$$\text{notes}_t \sim \text{Poisson}(\lambda_t)$$
$$\lambda_t = (1 - \text{cat}_t)\alpha + \text{cat}_t\beta$$
$$\text{cat}_t \sim \text{Bernoulli}(\kappa)$$
$$\kappa \sim \text{Beta}(4, 4)$$
$$\alpha \sim \text{Exponential}(1/10)$$
$$\beta \sim \text{Exponential}(1/10)$$

Stan code

```
generated quantities{
    vector[N] cat_impute;
    for ( i in 1:N ) {
        real logPxy;
        real logPy;
        if ( cat[i]==-1 ) {
            logPxy = log(kappa) +
                      poisson_lpmf( notes[i] | beta);
            logPy = log_mix( kappa ,
                    poisson_lpmf( notes[i] | beta ),
                    poisson_lpmf( notes[i] | alpha ) );
            cat_impute[i] = exp( logPxy - logPy );
        } else {
            cat_impute[i] = cat[i];
        }
    }//i
}
```

	Mean	StdDev	lower 0.89	upper 0.89	n_eff	Rhat
kappa	0.52	0.13	0.30	0.72	1000	1
beta	7.40	1.44	5.00	9.52	1000	1
alpha	17.48	2.49	13.61	21.43	1000	1
cat_impute[1]	**0.75**	**0.21**	**0.44**	**1.00**	**1000**	**1**
cat_impute[2]	0.00	0.00	0.00	0.00	1000	NaN
cat_impute[3]	1.00	0.00	1.00	1.00	1000	NaN
cat_impute[4]	**0.01**	**0.03**	**0.00**	**0.01**	**611**	**1**
cat_impute[5]	1.00	0.00	1.00	1.00	1000	NaN
cat_impute[6]	0.00	0.00	0.00	0.00	1000	NaN
cat_impute[7]	1.00	0.00	1.00	1.00	1000	NaN

https://gist.github.com/rmcelreath

Sly cats

- Cats are hard to detect! Birds always see them, but data logger misses them half the time.
- Unobserved cats as both "parameter" and "data"
- *Occupancy model*

$$
\begin{aligned}
\text{notes}_t &\sim \text{Poisson}(\lambda_t) \\
\lambda_t &= (1 - \text{cat}_t)\alpha + \text{cat}_t\beta \\
\text{cat}_{\text{obs},t} &\sim \text{Bernoulli}(\text{cat}_t \times \delta) \\
\text{cat}_t &\sim \text{Bernoulli}(\kappa) \\
\kappa &\sim \text{Beta}(4, 4) \\
\delta &\sim \text{Beta}(4, 4) \\
\alpha &\sim \text{Exponential}(1/10) \\
\beta &\sim \text{Exponential}(1/10)
\end{aligned}
$$

$$\text{notes}_t \sim \text{Poisson}(\lambda_t)$$
$$\lambda_t = (1 - \text{cat}_t)\alpha + \text{cat}_t\beta$$
$$\text{cat}_{\text{obs},t} \sim \text{Bernoulli}(\text{cat}_t \times \delta)$$
$$\text{cat}_t \sim \text{Bernoulli}(\kappa)$$
$$\kappa \sim \text{Beta}(4, 4)$$
$$\delta \sim \text{Beta}(4, 4)$$
$$\alpha \sim \text{Exponential}(1/10)$$
$$\beta \sim \text{Exponential}(1/10)$$

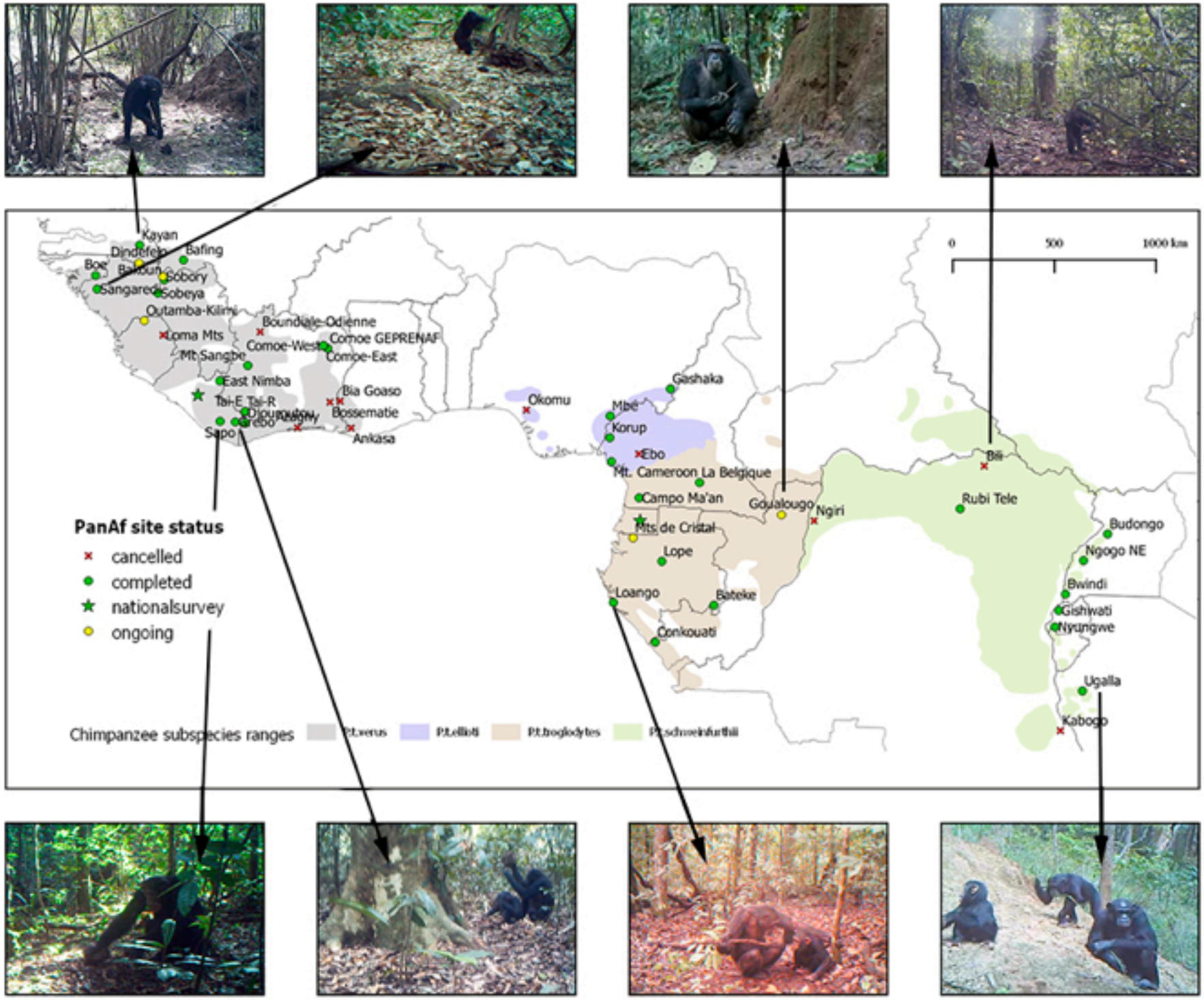
Stan code

```
model {
    beta ~ exponential( 0.1 );
    alpha ~ exponential( 0.1 );
    kappa ~ beta(4,4);
    delta ~ beta(4,4);

    for ( i in 1:N ) {
        if ( cat[i]==1 )
            // cat present and detected
            target += log(kappa) + log(delta) +
                        poisson_lpmf( notes[i] | beta );
        if ( cat[i]==0 ) {
            // cat not observed, but cannot be sure not there
            // marginalize over unknown cat state:
            // (1) cat present and not detected
            // (2) cat absent
            target += log_sum_exp(
                    log(kappa) + log1m(delta) +
                        poisson_lpmf( notes[i] | beta ),
                    log1m(kappa) +
                        poisson_lpmf( notes[i] | alpha ) );
        }//cat==0
    }//i
}
```

	Mean	StdDev	lower 0.89	upper 0.89	n_eff	Rhat
beta	7.70	1.42	5.30	9.74	1000	1
alpha	18.13	2.57	14.47	22.46	1000	1
kappa	**0.54**	**0.12**	**0.34**	**0.75**	**1000**	**1**
delta	**0.66**	**0.13**	**0.47**	**0.88**	**1000**	**1**

https://gist.github.com/rmcelreath

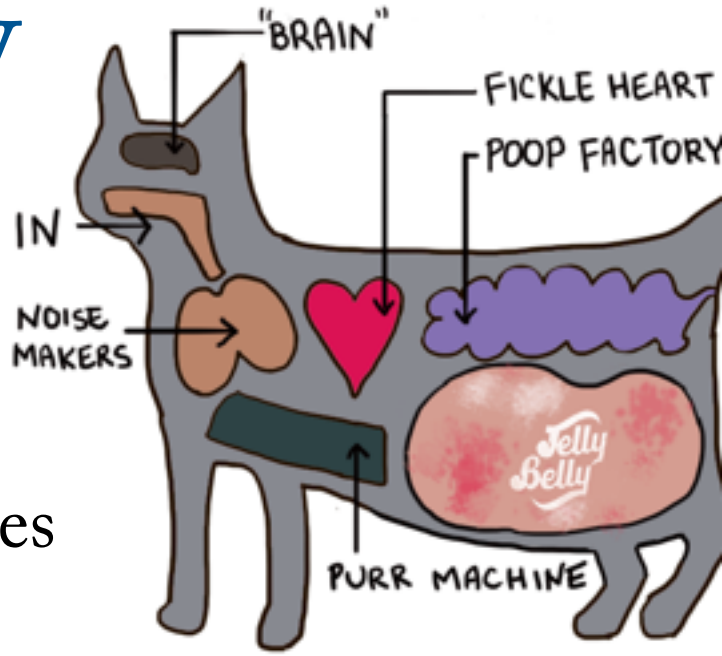


http://panafrican.eva.mpg.de/

Four Unifying Forces

- Unity of data/parameters, likelihoods/priors:
  1. Same derivations & calculations
  2. Same inferential force => e.g. shrinkage
  3. Do double duty, conditional on observation
  4. Can be both in same analysis

Benefits of insider view



- Not necessary, but useful
- Think scientifically, not statistically
  - Define generative model of all variables
  - Use observed variables in inference
- Direct solutions to common problems
  - Measurement messes, propagate uncertainty
  - But lots of computational challenges remain!
- Unified approach to construction
- Demystifying. Deflationary.
- Help in teaching — Bayes NOT likelihood + priors

A Modest Proposal

Convention	Proposal
Data	Observed variable
Parameter	Unobserved variable
Likelihood	Distribution
Prior	Distribution
Posterior	Conditional distribution
Estimate	*banished*
Random	*banished*