# Stan

## a Probabilistic Programming Language


*Core Development Team* *(in order of joining):*

Andrew Gelman, **Bob Carpenter**, Matt Hoffman, **Daniel Lee**, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, Allen Riddell, Marco Inacio, Jeffrey Arnold, Mitzi Morris, Rob Trangucci, Rob Goedman, Brian Lau, Jonah Sol Gabry, Alp Kucukelbir, Robert L. Grant, Dustin Tran, Alp Kucukelbir, Krzysztof Sakrejda, Aki Vehtari, Rayleigh Lei, Sebastian Weber



Stan 2.9.0 (April 2016) http://mc-stan.org


Stan

# Get the Slides

http://mc-stan.org/workshops/vanderbilt2016

## Example I

# Male Birth Ratio

# Birth Rate by Sex

- **Laplace**'s data on live births in Paris from 1745–1770:

| *sex* | *live births* |
|---|---|
| female | 241 945 |
| male | 251 527 |

- **Question 1** (Estimation)
  What is the birth rate of boys vs. girls?
- **Question 2** (Event Probability)
  Is a boy more likely to be born than a girl?
- Bayes (1763) set up the "Bayesian" model
- Laplace (1781, 1786) solved for the posterior

# Calculating Laplace's Answers

```
transformed data {
  int male;
  int female;
  male <- 251527;
  female <- 241945;
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  male ~ binomial(male + female, theta);
}
generated quantities {
  int<lower=0, upper=1> theta_gt_half;
  theta_gt_half <- (theta > 0.5);
}
```

# And the Answer is...

```
> fit <- stan("laplace.stan", iter=100000);
> print(fit, probs=c(0.005, 0.995), digits=3)
```

| | mean | 0.5% | 99.5% |
|---|---|---|---|
| *theta* | *0.51* | *0.508* | *0.512* |
| *theta_gt_half* | *1.00* | *1.000* | *1.000* |

- Q1: $\theta$ is 99% certain to lie in $(0.508, 0.512)$
- Q2: Laplace "morally certain" boys more prevalent

# Parameter Estimates

- Estimate **probability** that a parameter value in interval, e.g.,

$$\Pr[\theta \in (0.508, 0.512) \mid y]$$

- Conditions on observed data $y$

- **Bayesian parameter estimates are probabilistic**

# Event Probabilities

- Random variable defined by **indicator** function

```
theta_gt_half <- (theta > 0.5);
```

- Indicators are random variables
  - with boolean (0 or 1) values
  - defined in terms of parameters (and data)
- For Laplace's problem, calculus shows

$$\Pr[\theta \leq 0.5 \mid y] \approx 10^{-42}$$

- **Event probabilities are expectations of indicators**

## Warmup Exercise I

# Sample Variation

# Repeated i.i.d. Trials

- Suppose we repeatedly generate a random outcome from among several potential outcomes
- Suppose the outcome chances are the same each time
    - i.e., outcomes are independent and identically distributed (i.i.d.)
- For example, spin a fair spinner (without cheating), such as one from *Family Cricket*.


Image source: http://replaycricket.com/2010/10/29/family-cricket/

# Repeated i.i.d. Binary Trials

- Suppose the outcome is binary and assigned to 0 or 1; e.g.,
  - 20% chance of outcome 1: *ball in play*
  - 80% chance of outcome 0: *ball* not *in play*
- Consider different numbers of bowls delivered.
- How will proportion of successes in sample differ?

# Simulating Repeated Binary Trials

- R Code: `rbinom(10, N, 0.3)`
  - **N = 10** trials (10% to 50% success rate)

    2 2 1 3 3 2 3 2 2 5
  - **N = 100** trials (27% to 34% success rate)

    29 34 27 31 25 31 27 29 32 26
  - **N = 1000** trials (29% to 32% success rate)

    291 297 289 322 305 296 294 297 314 292
  - **N = 10,000** trials (29.5% to 30.7% success rate)

    3014 3031 3017 2886 2995 2944 3067 3069 3051 3068
- Deviation goes down at rate: $\mathcal{O}(1/\sqrt{N})$

# Simple Point Estimation

- Estimate chance of success $\theta$ by proportion of successes:

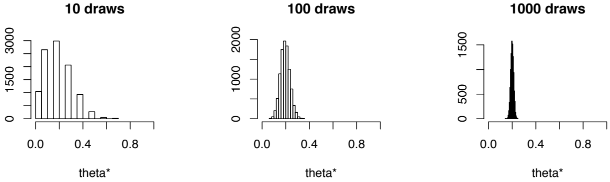
$$\theta^* = \frac{\text{successes}}{\text{attempts}}$$

- Simulation shows accuracy depends on the amount of data.
- Statistical inference includes quantifying uncertainty.
- Bayesian statistics is about using uncertainty in inference.

# Estimation Uncertainty

- Simulation of estimate variation due to sampling
- *not* a Bayesian posterior

```
> num_sims <- 10000;    N <- 100;    theta <- 0.2;
> hist(rbinom(num_sims, N, theta) / N,
       main=sprintf("%d draws",N), xlab="theta*");
```



# Estimator Bias

- **Bias:** expected difference of estimate ($\hat{\theta}$) from true value ($\theta$)

$$\text{bias} = \mathbb{E}[\theta - \hat{\theta}]$$

- Continuing previous example

```
> sims <- rbinom(10000, 1000, 0.2) / 1000
> mean(sims)
[1] 0.2002536
```

- Value of 0.2 is estimate of expectation
- Shows this estimator is *unbiased*

# Simple Point Estimation (cont.)

- **Central Limit Theorem:** *expected* error in $\theta^*$ goes down as

$$\frac{1}{\sqrt{N}}$$

- Each decimal place of accuracy requires $100\times$ more draws.
- Width of confidence intervals shrinks at the same rate.

- Can also use theory to show this estimator is unbiased.

# Pop Quiz! Cancer Clusters

- Why do lowest and highest cancer clusters look so similar?

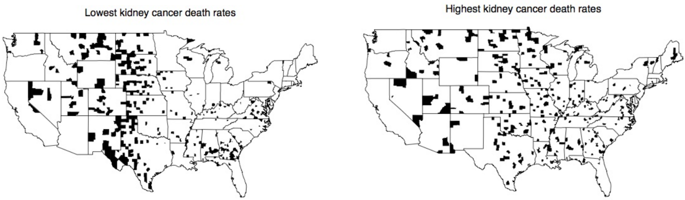



Image from Gelman et al., *Bayesian Data Analysis, 3rd Edition* (2013)

# Pop Quiz Answer

- Hint: mix earlier simulations of repeated i.i.d. trials with 20% success and sort:

| | | | | |
|---|---|---|---|---|
| 1/10 | 1/10 | 1/10 | 15/100 | 16/100 |
| 17/100 | 175/1000 | 179/1000 | 18/100 | 181/1000 |
| 188/1000 | 194/1000 | 198/1000 | 2/10 | 2/10 |
| 2/10 | 2/10 | 21/100 | 21/100 | 21/100 |
| 212/1000 | 213/1000 | 216/1000 | 223/1000 | 23/100 |
| 26/100 | 26/100 | 3/10 | 4/10 | 5/10 |

- More variation in observed rates with smaller sample sizes

- *Answer*: High cancer and low cancer counties are small populations

## Stan Example

# Repeated Binary Trials

# R: Simulate Data

- Generate data

```
> theta <- 0.30;
> N <- 20;
> y <- rbinom(N, 1, 0.3);

> y

 [1] 1 1 1 1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 1
```

- Calculate MLE as sample mean from data

```
> sum(y) / N

[1] 0.4
```

# RStan: Fit

```
> library(rstan);

> fit <- stan("bern.stan",
              data = list(y = y, N = N));

> print(fit, probs=c(0.1, 0.9));

Inference for Stan model: bern.
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  10%  90% n_eff Rhat
theta 0.41    0.00 0.10 0.28 0.55  1580    1
```

# Plug in Posterior Draws

- Extracting the posterior draws

```
> theta_draws <- extract(fit)$theta;
```

- Calculating posterior mean (estimator)

```
> mean(theta_draws);

[1] 0.4128373
```

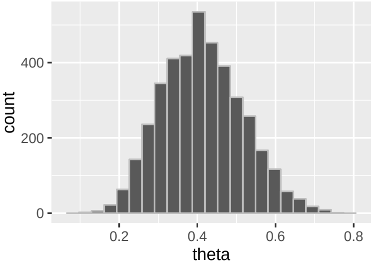
- Calculating posterior intervals

```
> quantile(theta_draws, probs=c(0.10, 0.90));

      10%       90%
0.2830349 0.5496858
```

# ggplot2: Plotting

```
theta_draws_df <- data.frame(list(theta = theta_draws));
plot <-
  ggplot(theta_draws_df, aes(x = theta)) +
  geom_histogram(bins=20, color = "gray");
plot;
```



## Warmup Exercise II

# Maximum Likelihood Estimation

# Observations, Counterfactuals, and Random Variables

- Assume we observe data $y = y_1, \ldots, y_N$
- Statistical modeling assumes even though $y$ is observed, the values could have been different
- John Stuart Mill first characterized this **counterfactual** nature of statistical modeling in:

  *A System of Logic, Ratiocinative and Inductive* (1843)
- In measure-theoretic language, $y$ is a **random variable**

# Likelihood Functions

- A **likelihood function** is a probability function (density, mass, or mixed)

$$p(y|\theta, x),$$

where

  - $\theta$ is a vector of **parameters**,
  - $x$ is some fixed **unmodeled data** (e.g., regression predictors or “features”),
  - $y$ is some fixed **modeled data** (e.g., observations)

- considered as a function $\mathcal{L}(\theta)$ of $\theta$ for fixed $x$ and $y$.

- can think of as a generative process for $y$how data $y$ is generated

# Maximum Likelihood Estimation

- **Estimate** parameters $\theta$ given observations $y$.
- Maximum likelihood estimation (MLE) chooses estimate that maximizes the likelihood function, i.e.,

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta p(y|\theta, x)$$

- This function of $\mathcal{L}$ and $y$ (and $x$) is called an **estimator**

# Example of MLE

- The frequency-based estimate

$$\theta^* = \frac{1}{N} \sum_{n=1}^{N} y_n,$$

is the observed rate of “success” (outcome 1) observations.

- This is the MLE for the model

$$p(y|\theta) \;=\; \prod_{n=1}^{N} p(y_n|\theta) \;=\; \prod_{n=1}^{N} \text{Bernoulli}(y_n|\theta)$$

where for $u \in \{0, 1\}$,

$$\text{Bernoulli}(u|\theta) = \begin{cases} \theta & \text{if } u = 1 \\ 1 - \theta & \text{if } u = 0 \end{cases}$$

# Example of MLE (cont.)

- First modeling *assumption* is that data are i.i.d.,

$$p(y|\theta) = \prod_{n=1}^{N} p(y_n|\theta)$$

- Second modeling *assumption* is form of likelihood,

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta)$$

# Example of MLE (cont.)

- The frequency-based estimate is the MLE
- First derivative is zero (indicating min or max),

$$\mathcal{L}_y'(\theta^*) = 0,$$

- Second derivative is negative (indicating max),

$$\mathcal{L}_y''(\theta^*) < 0.$$

# MLEs can be Dangerous!

- Recall the cancer cluster example
- Accuracy is low with small counts
- What we need are hierarchical models (stay tuned)

# Part I
# Bayesian Inference

# Bayesian Data Analysis

- "By Bayesian data analysis, we mean practical methods for making inferences from data using probability models for quantities we observe and about which we wish to learn."
- "The essential characteristic of Bayesian methods is their **explict use of probability for quantifying uncertainty** in inferences based on statistical analysis."

Gelman et al., *Bayesian Data Analysis*, 3rd edition, 2013

# Bayesian Methodology

- Set up **full probability model**
  - for all observable & unobservable quantities
  - consistent w. problem knowledge & data collection
- **Condition** on observed data
  - to caclulate posterior probability of unobserved quantities (e.g., parameters, predictions, missing data)
- **Evaluate**
  - model fit and implications of posterior
- **Repeat** as necessary

*Ibid.*

# Where do Models Come from?

- Sometimes model comes first, based on substantive considerations
  - toxicology, economics, ecology, . . .
- Sometimes model chosen based on data collection
  - traditional statistics of surveys and experiments
- Other times the data comes first
  - observational studies, meta-analysis, . . .
- Usually its a mix

# (Donald) Rubin's Philosophy

- All statistics is inference about missing data
- Question 1: What would you do if you had all the data?
- Question 2: What were you doing before you had any data?

(as relayed in course notes by Andrew Gelman)

# Model Checking

- Do the inferences make sense?
  - are parameter values consistent with model's prior?
  - does simulating from parameter values produce reasoable fake data?
  - are marginal predictions consistent with the data?
- Do predictions and event probabilities for new data make sense?
- ***Not***: Is the model true?
- ***Not***: What is Pr[model is true]?
- ***Not***: Can we "reject" the model?

# Model Improvement

- Expanding the model
  - hierarchical and multilevel structure ...
  - more flexible distributions (overdispersion, covariance)
  - more structure (geospatial, time series)
  - more modeling of measurement methods and errors
  - ...
- Including more data
  - breadth (more predictors or kinds of observations)
  - depth (more observations)

# Using Bayesian Inference

- Finds parameters consistent with prior info and data*
  - * if such agreement is possible
- Automatically includes uncertainty and variability
- Inferences can be plugged in directly
  - risk assesment
  - decision analysis

# Notation for Basic Quantities

- **Basic Quantities**
  - $y$: observed data
  - $\theta$: parameters (and other unobserved quantities)
  - $x$: constants, predictors for conditional (aka "discriminative") models
- **Basic Predictive Quantities**
  - $\tilde{y}$: unknown, potentially observable quantities
  - $\tilde{x}$: constants, predictors for unknown quantities

# Naming Conventions

- **Joint**: $p(y, \theta)$
- **Sampling / Likelihood**: $p(y|\theta)$
  - Sampling is function of $y$ with $\theta$ fixed (prob function)
  - Likelihood is function of $\theta$ with $y$ fixed (*not* prob function)
- **Prior**: $p(\theta)$
- **Posterior**: $p(\theta|y)$
- **Data Marginal (Evidence)**: $p(y)$
- **Posterior Predictive**: $p(\tilde{y}|y)$

# Bayes's Rule for Posterior

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y,\theta)}{p(y)} && \text{[def of conditional]} \\
&= \frac{p(y|\theta)\, p(\theta)}{p(y)} && \text{[chain rule]} \\
&= \frac{p(y|\theta)\, p(\theta)}{\int_\Theta p(y,\theta')\, d\theta'} && \text{[law of total prob]} \\
&= \frac{p(y|\theta)\, p(\theta)}{\int_\Theta p(y|\theta')\, p(\theta')\, d\theta'} && \text{[chain rule]}
\end{aligned}
$$

- *Inversion:* Final result depends only on sampling distribution (likelihood) $p(y|\theta)$ and prior $p(\theta)$

# Bayes's Rule up to Proportion

- If data $y$ is fixed, then

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)\,p(\theta)}{p(y)} \\
&\propto p(y|\theta)\,p(\theta) \\
&= p(y,\theta)
\end{aligned}
$$

- Posterior proportional to likelihood times prior
- Equivalently, posterior proportional to joint
- The nasty integral for data marginal $p(y)$ goes away

# Posterior Predictive Distribution

- Predict new data $\tilde{y}$ based on observed data $y$
- Marginalize out parameters from posterior

$$p(\tilde{y}|y) \;=\; \int_{\Theta} p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta.$$

- Averages predictions $p(\tilde{y}|\theta)$, weight by posterior $p(\theta|y)$
  - $\Theta = \{\theta \mid p(\theta|y) > 0\}$ is support of $p(\theta|y)$
- Allows continuous, discrete, or mixed parameters
  - integral notation shorthand for sums and/or integrals

# Event Probabilities

- Recall that an event $A$ is a collection of outcomes
- Suppose event $A$ is determined by indicator on parameters

$$f(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{if } \theta \notin A \end{cases}$$

- e.g., $f(\theta) = \mathrm{I}(\theta_1 > \theta_2)$ for $\Pr[\theta_1 > \theta_2 \mid y]$
- Bayesian event probabilities calculate posterior mass

$$\Pr[A] = \int_{\Theta} f(\theta)\, p(\theta|y)\, d\theta.$$

- Not frequentist, because involves parameter probabilities

## Example I

# Male Birth Ratio

# Laplace's Data and Problems

- Laplace's data on live births in Paris from 1745–1770:

| *sex* | *live births* |
|---|---|
| female | 241 945 |
| male | 251 527 |

- Question 1 (Event Probability)
  Is a boy more likely to be born than a girl?
- Question 2 (Estimate)
  What is the birth rate of boys vs. girls?
- Bayes formulated the basic binomial model
- Laplace solved the integral

# Binomial Distribution

- Binomial distribution is number of successes $y$ in $N$ i.i.d. Bernoulli trials with chance of success $\theta$
- If $y_1, \ldots, y_N \sim \text{Bernoulli}(\theta)$,

  then $(y_1 + \cdots + y_N) \sim \text{Binomial}(N, \theta)$
- The analytic form is

$$\text{Binomial}(y|N, \theta) \;=\; \binom{N}{y} \theta^y (1-\theta)^{N-y}$$

  where the binomial coefficient normalizes for permutations (i.e., which subset of $n$ has $y_n = 1$),

$$\binom{N}{y} = \frac{N!}{y!\,(N-y)!}$$

# Bayes's Binomial Model

- Data
  - $y$: total number of male live births (data: 241 945)
  - $N$ : total number of live births (data: 493 472)
- Parameter
  - $\theta \in (0, 1)$: proportion of male live births
- Likelihood

$$p(y|N,\theta) \;=\; \text{Binomial}(y|N,\theta) \;=\; \binom{N}{y}\theta^{y}(1-\theta)^{N-y}$$

- Prior

$$p(\theta) \;=\; \text{Uniform}(\theta \mid 0, 1) \;=\; 1$$

# Detour: Beta Distribution

- For parameters $\alpha, \beta > 0$ and $\theta \in (0,1)$,

$$\text{Beta}(\theta|\alpha,\beta) \;\propto\; \theta^{\alpha-1}\,(1-\theta)^{\beta-1}$$

- Euler's Beta function is used to normalize,

$$\text{B}(\alpha,\beta) \;=\; \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du \;=\; \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

so that

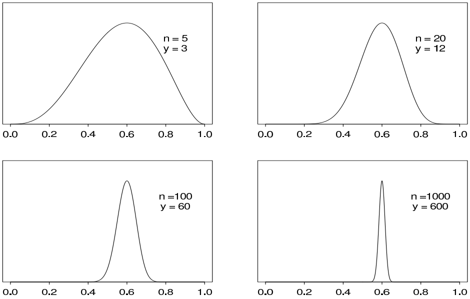
$$\text{Beta}(\theta|\alpha,\beta) \;=\; \frac{1}{\text{B}(\alpha,\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}$$

- Note: $\text{Beta}(\theta|1,1) = \text{Uniform}(\theta|0,1)$

- Note: $\Gamma()$ is continuous generalization of factorial

# Beta Distribution — Examples

- Unnormalized posterior density assuming uniform prior and $y$ successes out of $n$ trials (all with mean 0.6).




Gelman et al. (2013) *Bayesian Data Analysis*, 3rd Edition.

# Laplace Turns the Crank

- From Bayes's rule, the posterior is

$$p(\theta|y,N) \;=\; \frac{\mathsf{Binomial}(y|N,\theta)\,\mathsf{Uniform}(\theta|0,1)}{\int_{\Theta}\mathsf{Binomial}(y|N,\theta')\,p(\theta')\;d\theta'}$$

- Laplace calculated the posterior analytically

$$p(\theta|y,N) \;=\; \mathsf{Beta}(\theta\,|\,y+1,\;N-y+1).$$

# Estimation

- Posterior is Beta$(\theta \mid 1 + 241\,945,\ 1 + 251\,527)$
- Posterior mean:

$$\frac{1 + 241\,945}{1 + 241\,945 + 1 + 251\,527} \approx 0.490291\mathbf{3}$$

- Maximum likelihood estimate same as posterior mode (because of uniform prior)

$$\frac{241\,945}{241\,945 + 251\,527} \approx 0.490291\mathbf{2}$$

- As number of observations approaches $\infty$,
  MLE approaches posterior mean

# Event Probability Inference

- What is probability that a male live birth is more likely than a female live birth?

$$\begin{aligned} \mathrm{Pr}[\theta > 0.5] &= \int_{\Theta} \mathrm{I}[\theta > 0.5]\, p(\theta|y,N)\, d\theta \\ &= \int_{0.5}^{1} p(\theta|y,N)\, d\theta \\ &= 1 - F_{\theta|y,N}(0.5) \\ &\approx 10^{-42} \end{aligned}$$

- $\mathrm{I}[\phi] = 1$ if condition $\phi$ is true and 0 otherwise.
- $F_{\theta|y,N}$ is posterior cumulative distribution function (cdf).

# Mathematics vs. Simulation

- Luckily, we don't have to be as good at math as Laplace
- Nowadays, we calculate all these integrals by computer using tools like Stan

> If you wanted to do foundational research in statistics in the mid-twentieth century, you had to be bit of a mathematician, whether you wanted to or not. . . . if you want to do statistical research at the turn of the twenty-first century, you have to be a computer programmer.
>
> —from Andrew's blog

# Bayesian "Fisher Exact Test"

- Suppose we observe the following data on handedness

| | *sinister* | *dexter* | TOTAL |
|---|---|---|---|
| *male* | 9 ($y_1$) | 43 | 52 ($N_1$) |
| *female* | 4 ($y_2$) | 44 | 48 ($N_2$) |

- Assume likelihoods Binomial$(y_k \mid N_k, \theta_k)$, uniform priors

- Are men more likely to be lefthanded?

$$\begin{aligned} \Pr[\theta_1 > \theta_2 \mid y, N] &= \int_\Theta \mathrm{I}[\theta_1 > \theta_2]\, p(\theta \mid y, N)\, d\theta \\ &\approx 0.91 \end{aligned}$$

- Directly interpretable result; *not* a frequentist procedure

# Bayesian "Fisher Exact Test"

- Suppose we observe the following data on handedness

| | *sinister* | *dexter* | TOTAL |
|---|---|---|---|
| *male* | 9 ($y_1$) | 43 | 52 ($N_1$) |
| *female* | 4 ($y_2$) | 44 | 48 ($N_2$) |

- Assume likelihoods Binomial$(y_k|N_k, \theta_k)$, uniform priors
- Are men more likely to be lefthanded?

$$\Pr[\theta_1 > \theta_2 \mid y, N] \quad = \quad \int_\Theta \mathrm{I}[\theta_1 > \theta_2]\, p(\theta|y, N)\, d\theta$$

- Directly interpretable result; *not* a frequentist procedure

# Stan Binomial Comparison

```
data {
  int y[2];
  int N[2];
}
parameters {
  vector<lower=0,upper=1> theta[2];
}
model {
  y ~ binomial(N, y);
}
generated quantities {
  real boys_minus_girls;
  int boys_gt_girls;
  boys_minus_girls <- theta[1] - theta[2];
  boys_gt_girls <- (theta[1] > theta[2]);
}
```

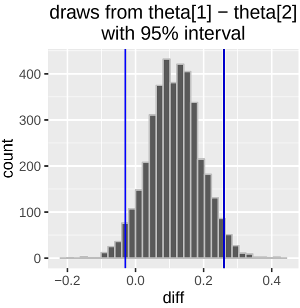
# Results

| | mean | 2.5% | 97.5% |
|---|---|---|---|
| theta[1] | 0.22 | 0.12 | 0.35 |
| theta[2] | 0.11 | 0.04 | 0.21 |
| boys_minus_girls | 0.12 | -0.03 | 0.26 |
| boys_gt_girls | 0.93 | 0.00 | 1.00 |

- $\Pr[\theta_1 > \theta_2 \mid y] \approx 0.93$
- $\Pr\left[(\theta_1 - \theta_2) \in (-0.03, 0.26) \mid y\right] = 95\%$

# Visualizing Posterior Difference

- Plot of posterior difference, $p(\theta_1 - \theta_2 \mid y, N)$ (men - women)




- Vertical bars: central 95% posterior interval $(-0.03, 0.26)$

## Technical Interlude

# Conjugate Priors

# Conjugate Priors

- Family $\mathcal{F}$ is a conjugate prior for family $\mathcal{G}$ if
  - prior in $\mathcal{F}$ and
  - likelihood in $\mathcal{G}$,
  - entails posterior in $\mathcal{F}$
- Before MCMC techniques became practical, Bayesian analysis mostly involved conjugate priors
- Still widely used because analytic solutions are more efficient than MCMC

# Beta is Conjugate to Binomial

- Prior: $p(\theta|\alpha,\beta) = \mathsf{Beta}(\theta|\alpha,\beta)$
- Likelihood: $p(y|N,\theta) = \mathsf{Binomial}(y|N,\theta)$
- Posterior:

$$
\begin{aligned}
p(\theta|y,N,\alpha,\beta) &\propto p(\theta|\alpha,\beta)\, p(y|N,\theta) \\
&= \mathsf{Beta}(\theta|\alpha,\beta)\, \mathsf{Binomial}(y|N,\theta) \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}\theta^{\alpha-1}\,(1-\theta)^{\beta-1}\ \binom{N}{y}\theta^{y}\,(1-\theta)^{N-y} \\
&\propto \theta^{y+\alpha-1}\,(1-\theta)^{N-y+\beta-1} \\
&\propto \mathsf{Beta}(\theta|\alpha+y,\ \beta+(N-y))
\end{aligned}
$$

# Chaining Updates

- Start with prior Beta$(\theta|\alpha, \beta)$
- Receive binomial data in $K$ statges $(y_1, N_1), \ldots, (y_K, N_K)$
- After $(y_1, N_1)$, posterior is Beta$(\theta|\alpha + y_1,\ \beta + N_1 - y_1)$
- Use as prior for $(y_2, N_2)$, with posterior

  Beta$(\theta|\alpha + y_1 + y_2,\ \ \beta + (N_1 - y_1) + (N_2 - y_2))$
- Lather, rinse, repeat, until final posterior

  Beta$(\theta|\alpha+y_1+\cdots+y_K,\ \ \beta+(N_1+\cdots+N_K)-(y_1+\cdots+y_K))$
- Same result as if we'd updated with combined data

  Beta$(y_1 + \cdots + y_K,\ \ N_1 + \cdots + N_K)$

# Part II
# (Un-)Bayesian Point Estimation

# MAP Estimator

- For a Bayesian model $p(y, \theta) = p(y|\theta)\, p(\theta)$, the max a posteriori (MAP) estimate maximizes the posterior,

$$\begin{aligned}
\theta^* &= \operatorname{arg\,max}_\theta\; p(\theta|y) \\
&= \operatorname{arg\,max}_\theta\; \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= \operatorname{arg\,max}_\theta\; p(y|\theta)p(\theta). \\
&= \operatorname{arg\,max}_\theta\; \log p(y|\theta) + \log p(\theta).
\end{aligned}$$

- *not* Bayesian because it doesn't integrate over uncertainty
- *not* frequentist because of distributions over parameters

# MAP and the MLE

- MAP estimate reduces to the MLE if the prior is uniform, i.e.,

$$p(\theta) = c$$

because

$$\begin{aligned}
\theta^* &= \arg\max_\theta \, p(y|\theta)\, p(\theta) \\
&= \arg\max_\theta \, p(y|\theta)\, c \\
&= \arg\max_\theta \, p(y|\theta).
\end{aligned}$$

# Penalized Maximum Likelihood

- The MAP estimate can be made palatable to frequentists via philosophical sleight of hand
- Treat the negative log prior $-\log p(\theta)$ as a "penalty"
- e.g., a Normal$(\theta|\mu,\sigma)$ prior becomes a penalty function

$$\lambda_{\theta,\mu,\sigma} \;=\; -\left(\log\sigma + \frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2\right)$$

- Maximize sum of log likelihood and negative penalty

$$\begin{aligned} \theta^* &= \operatorname{arg\,max}_\theta \; \log p(y|\theta,x) - \lambda_{\theta,\mu,\sigma} \\ &= \operatorname{arg\,max}_\theta \; \log p(y|\theta,x) + \log p(\theta|\mu,\sigma) \end{aligned}$$

# Proper Bayesian Point Estimates

- Choose estimate to minimize some loss function
- To minimize expected squared error (L2 loss), $\mathbb{E}[(\theta - \theta')^2 \mid y]$, use the posterior mean

$$\hat{\theta} = \arg\min_{\theta'} \mathbb{E}[(\theta - \theta')^2 \mid y] = \int_{\Theta} \theta \times p(\theta|y)\, d\theta.$$

- To minimize expected absolute error (L1 loss), $\mathbb{E}[|\theta - \theta'|]$, use the posterior median.
- Other loss (utility) functions possible, the study of which falls under decision theory
- All share property of involving full Bayesian inference.

# Point Estimates for Inference?

- Common in machine learning to generate a point estimate $\theta^*$, then use it for inference, $p(\tilde{y}|\theta^*)$
- This is **defective** because it

  **underestimates uncertainty**.

- To properly estimate uncertainty, apply full Bayes
- A major focus of statistics and decision theory is estimating uncertainty in our inferences

## Philosophical Interlude

# What is Statistics?

# Exchangeability

- Roughly, an exchangeable probability function is such that for a sequence of random variables $y = y_1, \ldots, y_N$,

$$p(y) = p(\pi(y))$$

  for every $N$-permutation $\pi$ (i.e, a one-to-one mapping of $\{1, \ldots, N\}$)

- i.i.d. implies exchangeability, but not vice-versa

# Exchangeability

- Roughly, an exchangeable probability function is such that for a sequence of random variables $y = y_1, \ldots, y_N$,

$$p(y) = p(\pi(y))$$

  for every $N$-permutation $\pi$ (i.e, a one-to-one mapping of $\{1, \ldots, N\}$ to itself)

- i.i.d. implies exchangeability, but not vice-versa
  - $(y_1, y_2)$ exchangeable, but not independent in

$$(y_1, y_2) \sim \text{MultiNormal}\left(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

# Exchangeability Assumptions

- Models almost always make some kind of exchangeability assumption
- Typically when other knowledge is not available
  - e.g., treat voters as conditionally i.i.d. given their age, sex, income, education level, religous affiliation, and state of residence
  - But voters have many more properties (hair color, height, profession, employment status, marital status, car ownership, gun ownership, etc.)
  - Missing predictors introduce additional error (on top of measurement error)

# Random Parameters: Doxastic or Epistemic?

- Bayesians treat distributions over parameters as epistemic (i.e., about knowledge)
- They do *not* treat them as being doxastic (i.e., about beliefs)
- Priors encode our knowledge before seeing the data
- Posteriors encode our knowledge after seeing the data
- Bayes's rule provides the way to update our knowledge
- People like to pretend models are ontological (i.e., about what exists)

# Arbitrariness: Priors vs. Likelihood

- Bayesian analyses often criticized as subjective (arbitrary)
- Choosing priors is no more arbitrary than choosing a likelihood function (or an exchangeability/i.i.d. assumption)
- As George Box famously wrote (1987),

*"All models are wrong, but some are useful."*

- This does not just apply to Bayesian models!

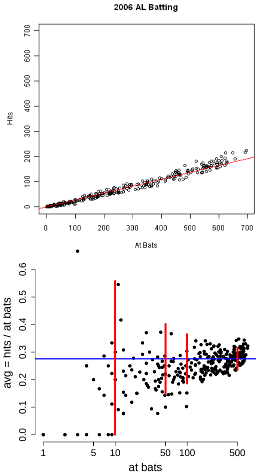
# Part IV
# Hierarchical Models

# Baseball At-Bats

- For example, consider baseball batting ability.
  - Baseball is sort of like cricket, but with round bats, a one-way field, stationary "bowlers", four bases, short games, and no draws
- Batters have a number of "at-bats" in a season, out of which they get a number of "hits" (hits are a good thing)
- Nobody with higher than 40% success rate since 1950s.
- No player (excluding "bowlers") bats much less than 20%.
- Same approach applies to hospital pediatric surgery complications (a BUGS example), reviews on Yelp, test scores in multiple classrooms, . . .

# Baseball Data

- Hits versus at bats for the 2006 American League season
- Not much variation in ability!
- Ignore skill vs. at-bats relation
- Note uncertainty of MLE



# Pooling Data

- How do we estimate the ability of a player who we observe getting 6 hits in 10 at-bats? Or 0 hits in 5 at-bats? Estimates of 60% or 0% are absurd!
- Same logic applies to players with 152 hits in 537 at bats.
- *No pooling*: estimate each player separately
- *Complete pooling*: estimate all players together (assume no difference in abilities)
- *Partial pooling*: somewhere in the middle
  - use information about other players (i.e., the population) to estimate a player's ability

# Hierarchical Models

- Hierarchical models are principled way of determining how much pooling to apply.
- Pull estimates toward the population mean based on amount of variation in population
  - low variance population: more pooling
  - high variance population: less pooling
- In limit
  - as variance goes to 0, get complete pooling
  - as variance goes to $\infty$, get no pooling

# Hierarchical Batting Ability

- Instead of fixed priors, estimate priors along with other parameters
- Still only uses data once for a single model fit
- Data: $y_n, B_n$: hits, at-bats for player $n$
- Parameters: $\theta_n$: ability for player $n$
- Hyperparameters: $\alpha, \beta$: population mean and variance
- Hyperpriors: fixed priors on $\alpha$ and $\beta$ (hardcoded)

# Hierarchical Batting Model (cont.)

$$\begin{aligned} y_n &\sim \text{Binomial}(B_n, \theta_n) \\ \theta_n &\sim \text{Beta}(\alpha, \beta) \\ \frac{\alpha}{\alpha+\beta} &\sim \text{Uniform}(0, 1) \\ (\alpha+\beta) &\sim \text{Pareto}(1.5) \end{aligned}$$

- Sampling notation syntactic sugar for:

  $p(y, \theta, \alpha, \beta) = \text{Pareto}(\alpha+\beta|1.5) \prod_{n=1}^{N} \left(\text{Binomial}(y_n|B_n, \theta_n)\ \text{Beta}(\theta_n|\alpha, \beta)\right)$

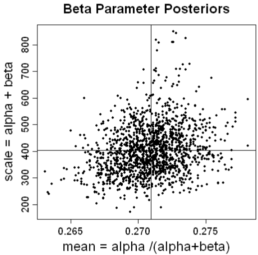
- Pareto provides power law: $\text{Pareto}(u|\alpha) \propto \frac{\alpha}{u^{\alpha+1}}$
- Should use more informative hyperpriors!

# Stan Program

```
data {
  int<lower=0> N;                // items
  int<lower=0> K[N];             // initial trials
  int<lower=0> y[N];             // initial successes
}
parameters {
  real<lower=0, upper=1> phi;         // population chance of suc
  real<lower=1> kappa;                // population concentration
  vector<lower=0, upper=1>[N] theta;  // chance of success
}
model {
  kappa ~ pareto(1, 1.5);                          // hyperprior
  theta ~ beta(phi * kappa, (1 - phi) * kappa);   // prior
  y ~ binomial(K, theta);                          // likelihood
}
```

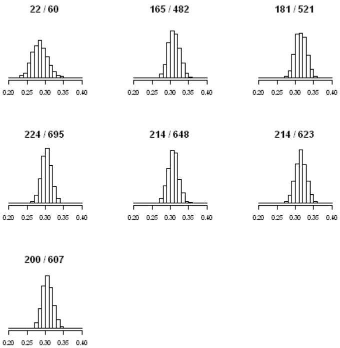
# Hierarchical Prior Posterior

- Draws from posterior (crosshairs at posterior mean)
- Prior population mean: 0.271
- Prior population scale: 400
- Together yield prior std dev of 0.022
- Mean is better estimated than scale (typical)



# Posterior Ability (High Avg Players)

- Histogram of posterior draws for high-average players
- Note uncertainty grows with lower at-bats



# Multiple Comparisons

- Who has the highest ability (based on this data)?
- Probabilty player $n$ is best is

| *Average* | *At-Bats* | Pr[best] |
|---|---|---|
| .347 | 521 | 0.12 |
| .343 | 623 | 0.11 |
| .342 | 482 | 0.08 |
| .330 | 648 | 0.04 |
| .330 | 607 | 0.04 |
| .367 | 60 | 0.02 |
| .322 | 695 | 0.02 |

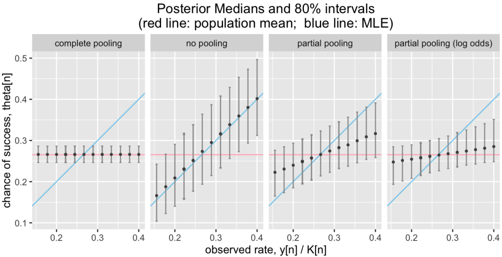
- No clear winner—sample size matters.
- In last game (of 162), Mauer (Minnesota) edged out Jeter (NY)

# Efron & Morris (1975) Data

| | FirstName | LastName | Hits | At.Bats | Rest.At.Bats | Rest.Hits |
|---|---|---|---|---|---|---|
| 1 | Roberto | Clemente | 18 | 45 | 367 | 127 |
| 2 | Frank | Robinson | 17 | 45 | 426 | 127 |
| 3 | Frank | Howard | 16 | 45 | 521 | 144 |
| 4 | Jay | Johnstone | 15 | 45 | 275 | 61 |
| 5 | Ken | Berry | 14 | 45 | 418 | 114 |
| 6 | Jim | Spencer | 14 | 45 | 466 | 126 |
| 7 | Don | Kessinger | 13 | 45 | 586 | 155 |
| 8 | Luis | Alvarado | 12 | 45 | 138 | 29 |
| 9 | Ron | Santo | 11 | 45 | 510 | 137 |
| 10 | Ron | Swaboda | 11 | 45 | 200 | 46 |
| 11 | Rico | Petrocelli | 10 | 45 | 538 | 142 |

# Pooling Estimates

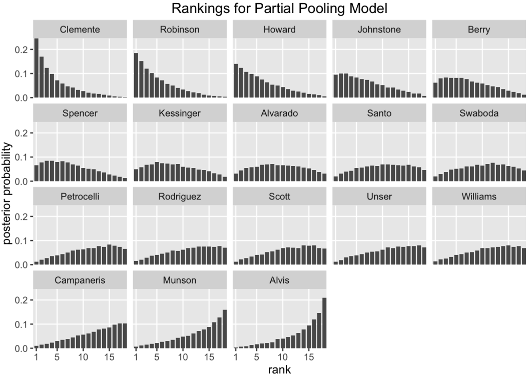
- Case Study: Repeated Binary Trials (mc-stan.org)



# Ranking

```
generated quantities {
  int<lower=1, upper=N> rnk[N];      // rank of player n
  {
    int dsc[N];
    dsc <- sort_indices_desc(theta);
    for (n in 1:N)
      rnk[dsc[n]] <- n;
  }
```

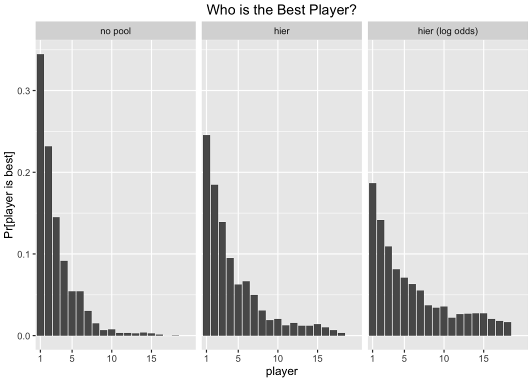
# Posterior Ranks



# Who is Best? Stan Code

```
generated quantities {
  ...
  int<lower=0, upper=1> is_best[N];  // Pr[player n highest chanc
  ...
  for (n in 1:N)
    is_best[n] <- (rnk[n] == 1);
  ...
```

# Who is Best? Posterior

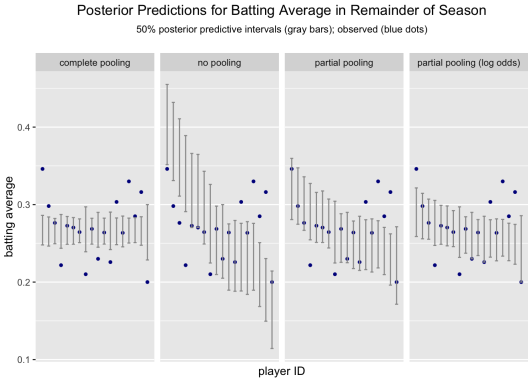


# Posterior Predictive Inference

- How do we predict new outcomes (e.g., rest of season)?

```
data {
  int<lower=0> K_new[N];       // new trials
  int<lower=0> y_new[N];       // new successes
  ...
generated quantities {
  int<lower=0> z[N];  // posterior prediction
  for (n in 1:N)
    z[n] <- binomial_rng(K_new[n], theta[n]);
```

# Posterior Predictions



# Posterior Predictive Check

- Replicate data from paraemters

```
generated quantities {
  ...
  for (n in 1:N)
    y_rep[n] <- binomial_rng(K[n], theta[n]);
  for (n in 1:N)
    y_pop_rep[n] <- binomial_rng(K[n],
                                 beta_rng(phi * kappa,
                                          (1 - phi) * kappa));
  min_y_rep <- min(y_rep);
  sd_y_rep <- sd(to_vector(y_rep));
  p_min <- (min_y_rep >= min_y);
  p_sd <- (sd_y_rep >= sd_y);
}
```

# Posterior $p$-Values