Bayesian Statistical Methods for Astronomy

Part I: Foundations

David A. van Dyk

Statistics Section, Imperial College London

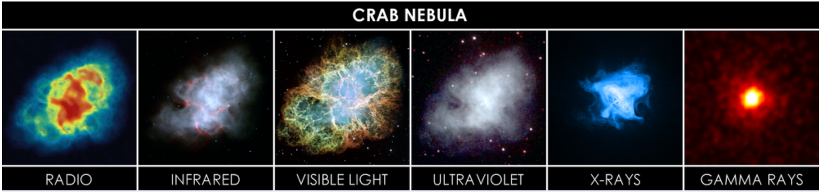
Center for Astrophysics | Harvard & Smithsonian
February 2025

Introduction

Massive new data streams are opening up a world of opportunities for data scientists!

1. Astrostatistics: Improved data quality and quantity lead to more interesting statistical models
2. Data-driven versus Science-driven methods
3. Predictive models versus Descriptive models
4. Tradeoff: computational speed and statistical principles
5. These issues are not unique to astronomy!

Astronomy: Dramatic Increase in Data Quality & Quantity



The crab nebula in radio, infrared, visible, ultraviolet, x-ray and gamma-ray wavelengths.

Sources: Radio: NRAO/AUI and M. Bietenholz, J.M. Uson, T.J. Cornwell; Infrared: NASA/JPL-Caltech/R. Gehrz (University of Minnesota); Visible: NASA, ESA, J. Hester and A.Loll (Arizona State University); Ultraviolet: NASA/Swift/E. Hoversten, PSU, X-ray: NASA/CXC/SAO/F. Seward et al.; Gamma: NASA/DOE/Fermi LAT/R. Buehler

- high resolution spectrography and imaging across the electromagnetic spectrum,
- different instruments/data and characteristics require tailored methods,
- space-based telescopes tailored to specific scientific goals,
- data are *not just massive:* they are rich, deep, & complex.

Event Lists in High-Energy Astrophysics

X-ray and γ-ray astronomers record photons (aka events):

- two-dimensional sky coordinates
- energy (e.g., color)
- time of arrival

Conceptualize as a 4-way table or data cloud

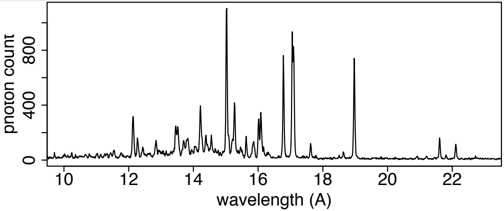
- High-energy (X-ray, γ-ray): few counts – Poisson models
- Low energy (optical): many counts – Gaussian models

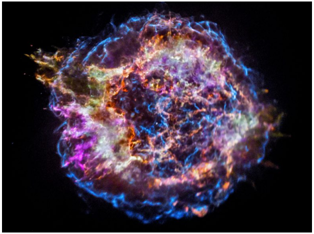
Typically focus on one of the marginal tables.

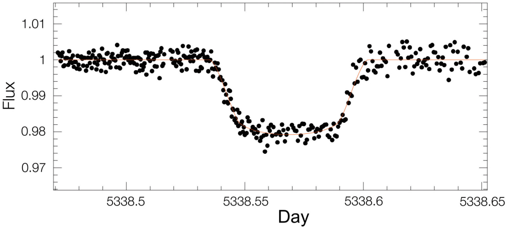
- Spectra, Image or Time Series

Ultimately we wish to combine information across the electromagnetic spectrum

Typically Focus on One of the Marginals



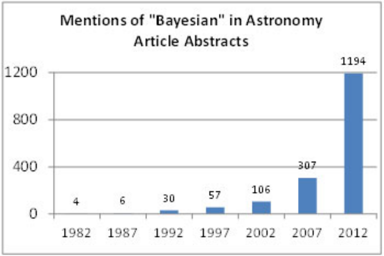


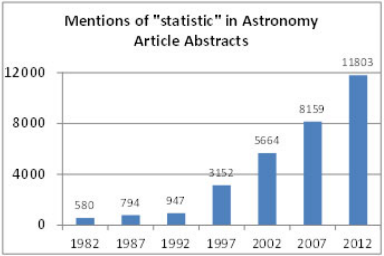


[Capella Spectrum; Img Credit: NASA/CXC/SAO; Exoplanet LC (WASP-19b), Credit: TRAPPIST/M Gillon/ESO]

Bayesian Renaissance in Astronomy

The use of Statistical Methods in general and Bayesian Methods in particular has grown exponentially in Astronomy.





Source: `http://magazine.amstat.org/blog/2013/12/01/science-policy-intel/`

Why Use Bayesian Methods?

Advantages of Bayesian methods:

- Directly model complexities of sources and instruments.
- Allows science-driven modeling. *(Not just predictive modeling.)*
- Combine multiple information sources and/or data streams.
- Allow hierarchical or multi-level structures in data/models.
- Bayesian methods have clear mathematical foundations and can be used to derive principled statistical methods.
- Sophisticated computational methods available.

Challenges:

- Require us to specify “prior distributions” on unknown model parameters.

Outline of Topics

1. BACKGROUND: Motivation; modern Bayesian tools; comparisons with likelihood methods; evaluating an estimator.
2. BASIC MODELS: Poisson, binomial, and normal models; conjugate, informative, non-informative, and Jeffries prior distributions; summarizing posterior inference; the posterior as an average of the prior and data; nuisance parameters.
3. MODEL FITTING: (Markov chain) Monte Carlo Methods, convergence detection, data augmentation
4. HIERARCHICAL MODELS: Random-effects models and shrinkage; Multilevel models; Examples: selection effects, spectral and image analysis in high-energy astrophysics.
5. MODEL CHECKING, SELECTION, AND IMPROVEMENT: Posterior predictive checks, Bayes factors, comparisons with significance tests and p-values.

Outline

Outline

1. Foundations of Bayesian Data Analysis
   - Probability
   - Bayesian Analysis of Standard Poisson Model
   - Building Blocks of Modern Bayesian Analyses

2. Further Topics with Univariate Parameter Models
   - Bayesian Analysis of Standard Binomial Model
   - Transformations
   - Prior Distributions
   - Comparisons with Frequency Based Methods

Defining Probability

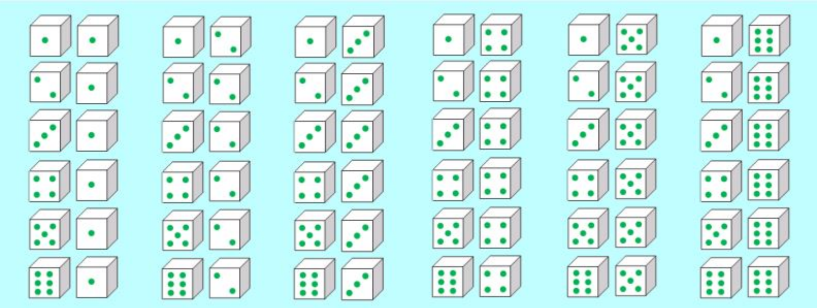
What do we mean by:

- $\mathsf{Pr}(\text{Roll two dice and get doubles}) =$
- $\mathsf{Pr}(\text{Rain today}) =$
- $\mathsf{Pr}\,(\text{catch a train departing South Station in 40 minutes}) =$
- $\pi(T) = \mathsf{Pr}\,(\text{catch train leaving in 40 min if I leave at time } T) =$

How should we define "probability"?

Rolling Dice

Suppose we roll two dice:



- Let $\mathcal{S}$ be the set of possible outcomes.

Mathematical Definition of Probability

Definition

(Kolmogorov Axioms) A <u>*probability function*</u> *is a function such that*

i) $\Pr(A) \geq 0$, *for all subsets of* $\mathcal{S}$.

ii) $\Pr(\mathcal{S}) = 1$.

iii) *For any pair of disjoint subsets,* A_1 *and* A_2*, of* $\mathcal{S}$*,*
$\Pr(A_1 \text{ or } A_2) = Pr(A_1) + Pr(A_2)$.[a]

[a] (Countable additivity) More generally, if $A_1, A_2, \ldots$ are pairwise disjoint subsets of $\mathcal{S}$ then $\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$.

But what does this this mean in real applications? How do we interpret a probability?

Defining Probability

What do we mean by:

- $\Pr(\text{Roll two dice and get doubles}) =$
- $\Pr(\text{Rain today}) =$
- Pr (catch a train departing King's Cross in 40 minutes) $=$
- $\pi(T) =$ Pr (catch train leaving in 40 min if I leave at time T) $=$

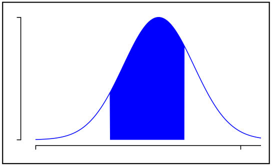
How should we define "probability"?

- Frequency-based definition.
- Subjective definition.
- Advantages and Difficulties of each.
- Is there a right or a wrong definition?

The Calculus of Probability

I assume you are familiar with:

- Probability density and mass functions, e.g.,
  - $\Pr(a < X < b) = \int_a^b p_X(x)dx$ or $\Pr(a \leq X \leq b) = \sum_{x=a}^b p_X(x)$
  - $\int_{-\infty}^{\infty} p(x)dx = 1$
- Joint probability functions, e.g.,
  - $\Pr(a < X < b \text{ and } Y > c) = \int_a^b \int_c^\infty p_{XY}(x,y)dydx$
  - $p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)dy$
- Conditional probability functions, e.g.,
  - $p_Y(y|x) = p_{XY}(x,y)/p_X(x)$
  - $p_{XY}(x,y) = p_X(x)p_Y(y|x)$



When it is clear from context, we omit the subscripts: $p(x) = p_X(x)$.

Bayes Theorem

Bayes Theorem allows us to reverse a conditional probability:

Theorem

Bayes Theorem:

$$p_Y(y|x) = \frac{p_X(x|y)p_Y(y)}{p_X(x)} \propto p_X(x|y)p_Y(y)$$

- Bayes Theorem follows from applying the definition of conditional probability twice:

$$p_Y(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} = \frac{p_X(x|y)p_Y(y)}{p_X(x)} \propto p_X(x|y)p_Y(y)$$

- The denominator does not depend on y and thus can be viewed as a normalizing constant. *Advantage?*

Outline

1. Foundations of Bayesian Data Analysis
   - Probability
   - Bayesian Analysis of Standard Poisson Model
   - Building Blocks of Modern Bayesian Analyses

2. Further Topics with Univariate Parameter Models
   - Bayesian Analysis of Standard Binomial Model
   - Transformations
   - Prior Distributions
   - Comparisons with Frequency Based Methods

A Poisson Model

Consider a Poisson model for a photon counting detector.

- Simplest case: single-bin detector

$$Y \stackrel{\text{dist}}{\sim} \textsc{Poisson}(\lambda_S \tau).$$

(τ is the observation time in seconds and λ_S is expected counts/sec.)

- The sampling distribution is the probability function of data:

$$p_Y(y|\lambda_S) = \frac{e^{-\lambda_S \tau}(\lambda_S \tau)^y}{y!}.$$

Definition

The likelihood function is the sampling distribution viewed as a function of the parameter. Constant factors may be omitted.

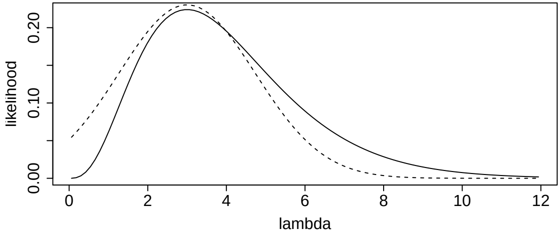
The maximum likelihood estimator (MLE) is the value of the parameter that maximizes the likelihood.

Likelihood for Poisson Model

Likelihood Function: For a single-bin detector,

$$\text{likelihood}(\lambda_S) = \frac{e^{-\lambda_S \tau}(\lambda_S \tau)^y}{y!} \qquad \text{loglikelihood}(\lambda_S) = -\lambda_S\tau + y\log(\lambda_S)$$

Maximum Likelihood Estimation: Suppose $y = 3$ with $\tau = 1$



The likelihood and its normal approximation.

MLE: $\hat{\lambda}_S = \dfrac{y}{\tau}$

Can estimate λ_S and its error bars.

Data-Appropriate Models and Methods

- Many methods based on χ^2 or Gaussian assumptions.
- Bayesian/Likelihood methods easily incorporate more appropriate distributions.
- E.g., for count data, we use a Poisson likelihood:

$$\chi^2 \text{ fitting:} \qquad -\sum_{\text{bins}} \frac{(y_i - \lambda_i)^2}{\sigma_i^2}$$

$$\text{Gaussian Loglikelihood:} \qquad -\sum_{\text{bins}} \sigma_i - \sum_{\text{bins}} \frac{(y_i - \lambda_i)^2}{\sigma_i^2}$$

$$\text{Poisson Loglikelihood:} \qquad -\sum_{\text{bins}} \lambda_i + \sum_{\text{bins}} y_i \log \lambda_i$$

A Prior Distribution for Poisson Model

Definition

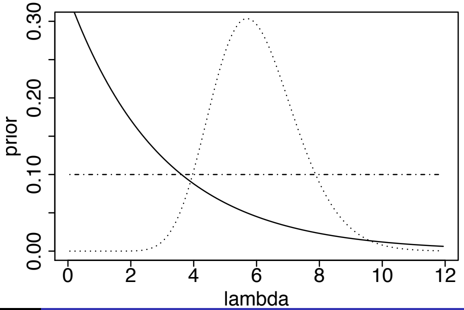
The prior distribution quantifies knowledge regarding parameters obtained prior to the current observation.

The *gamma distribution* is a flexible family of prior dist'ns:

$$p(\lambda_S) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_S^{\alpha-1} e^{-\beta\lambda_S}$$

for $\lambda_S > 0$.

- $\mathrm{E}(\lambda_S) = \alpha/\beta$
- $\mathrm{Var}(\lambda_S) = \alpha/\beta^2$



The Posterior Distribution for Poisson Model

Definition

The posterior distribution quantifies combined knowledge for parameters obtained prior to and with the current observation.

Bayes Theorem and the Posterior Distribution:

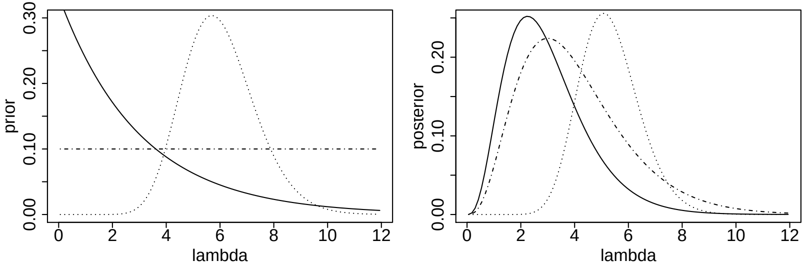
$$
\begin{aligned}
p(\lambda_S|y) &= p(y|\lambda_S)p(\lambda_S)/p(y) \\
\text{posterior}(\lambda_S|y) &\propto \text{likelihood}(\lambda_S|y) \times p(\lambda_S) \\
&\propto \frac{(\lambda_S\tau)^y e^{-\lambda_S\tau}}{y!} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_S^{\alpha-1}e^{-\beta\lambda_S} \\
&\propto \lambda_S^y e^{-\lambda_S\tau} \times \lambda_S^{\alpha-1}e^{-\beta\lambda_S} \\
&\propto \lambda_S^{y+\alpha+1}e^{-(\tau+\beta)\lambda_S}
\end{aligned}
$$

So:

$$\lambda_S|y \sim \textsc{Gamma}(y+\alpha, \beta+\tau)$$

The Posterior Distribution for Poisson Model

The posterior dist'n combines past and current information:



Bayesian analyses rely on probability theory.

Summary: Bayesian Analysis of Poisson Model

Definition

If the prior and the posterior distributions are of the same family, the prior dist'n is called that likelihood's conjugate prior distribution.

$$\boxed{\begin{array}{l}\text{If } Y|\lambda_S \stackrel{\text{dist}}{\sim} \textsc{Poisson}(\lambda_S\tau) \text{ and } \lambda_S \stackrel{\text{dist}}{\sim} \textsc{Gamma}(\alpha,\beta) \\ \text{then } \lambda_S|Y \stackrel{\text{dist}}{\sim} \textsc{Gamma}(y+\alpha, \tau+\beta).\end{array}}$$

- Conjugate prior distributions simplify computation!
- Using formulae for the Gamma distribution:
  - A Bayesian estimator of λ_S: $\mathrm{E}(\lambda_S|y) = \dfrac{y+\alpha}{\tau+\beta}$
  - A Bayesian error bar: $\sqrt{\mathrm{Var}(\lambda_S|Y)} = \dfrac{\sqrt{y+\alpha}}{\tau+\beta}$

"Prior Data"

Compare the MLE and the posterior expectation of λ_S:

$$\text{MLE}(\lambda_S) = \frac{y}{\tau} \qquad \mathrm{E}(\lambda_S|y) = \frac{y+\alpha}{\tau+\beta}$$

- The prior distribution has as much influence as α observed events in an exposure of β seconds.
- We can use this formulation of the prior in terms of "*prior data*" to
  - meaningfully specify the prior distribution for λ_S and
  - limit the influence of the prior distribution.

Outline

Model Specification

- The first step in a Bayesian analysis is specifying the statistical model
- This consists of specification of
  - the prior distribution
  - the likelihood function
- Both of these involves subjective choices
  - Comprehensive description can be overly complex.
  - Parsimony: simple w/out compromising scientific objectives.
  - What is a model?
  - What do we model? Or consider fixed?
    (E.g., calibration, preprocessing, selection, etc.)

All models are wrong, but some are useful.

—George Box

Multilevel (and Hierarchical) Models

Example: Background contamination in a single bin detector

- Contaminated source counts: $y = y_S + y_B$
- Background counts: x
- Background exposure is 24 times source exposure.

A Poisson Multi-Level Model:

LEVEL 1: $y|y_B, \lambda_S \stackrel{\text{dist}}{\sim} \text{Poisson}(\lambda_S) + y_B$,

LEVEL 2: $y_B|\lambda_B \stackrel{\text{dist}}{\sim} \text{Pois}(\lambda_B)$ and $x|\lambda_B \stackrel{\text{dist}}{\sim} \text{Pois}(\lambda_B \cdot 24)$,

LEVEL 3: specify a prior distribution for λ_B, λ_S.

Each level of the model specifies a dist'n given unobserved quantities whose dist'ns are given in lower levels.

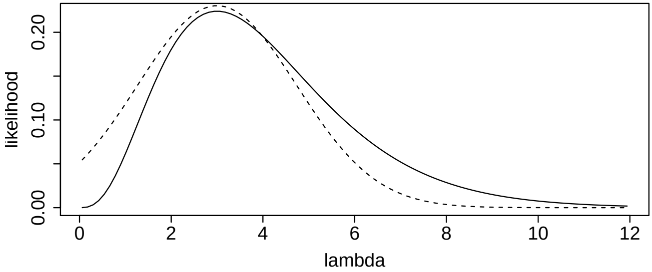
Bayesian Statistical Summaries

1. The full statistical summary: the posterior distribution.
2. But researchers would like summaries:

   A parameter estimate: The posterior mean.

   An error bar: The posterior standard deviation.

But is this enough??



Posterior Intervals or Regions

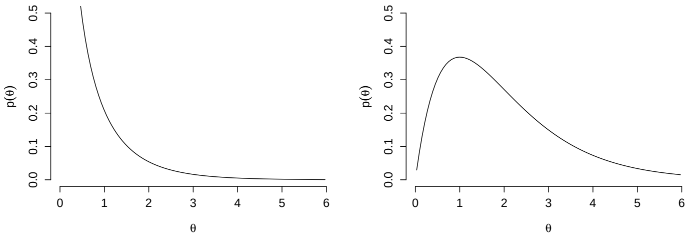
For non-Gaussian posterior dist'ns, we find L and U so that

$$\Pr(L < \theta < U|y) = \int_L^U p(\theta|y)d\theta = 68\% \text{ or } 95\% \text{ or } \ldots$$

or more generally, Θ so that

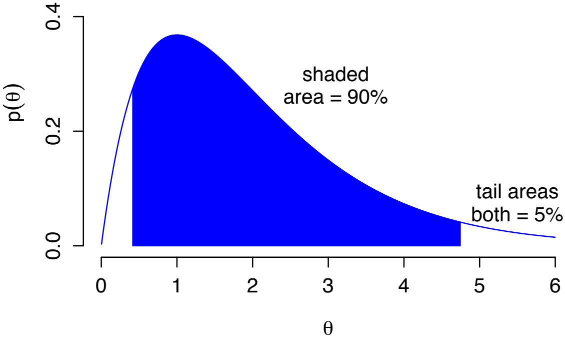
$$\Pr(\theta \in \Theta|y) = \int_{\theta\in\Theta} p(\theta|y)d\theta = 68\% \text{ or } 95\% \text{ or } \ldots$$

But the choice is not unique! Are there optimal choices?



Choice of Posterior Intervals

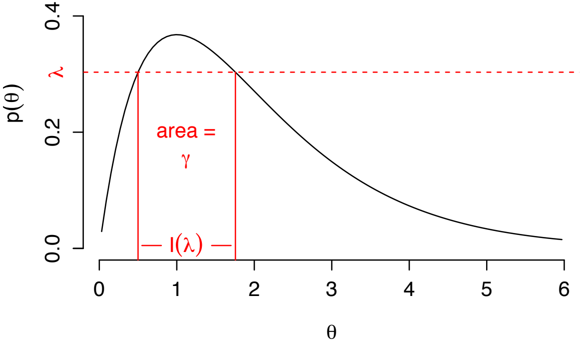
The Equal-Tailed Interval



- The simplest interval to compute (e.g., via Monte Carlo).
- Preserved under monotonic transformations.
  - E.g., If (L_θ, U_θ) is a 95% equal-tailed interval for θ, then $\left(\log(L_\theta), \log(U_\theta)\right)$ is a 95% equal-tailed interval for $\log(\theta)$

Choice of Posterior Intervals (con't)

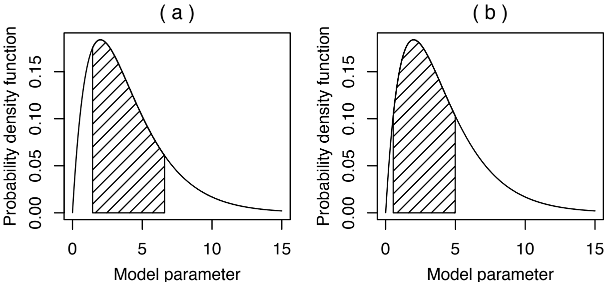
The Highest Posterior Density (HPD) Interval



- As λ decrease, probability (γ) of interval ($I(\lambda)$) increases.
- HPD interval is shortest interval of a given probability.

Choice of Posterior Intervals (con't)

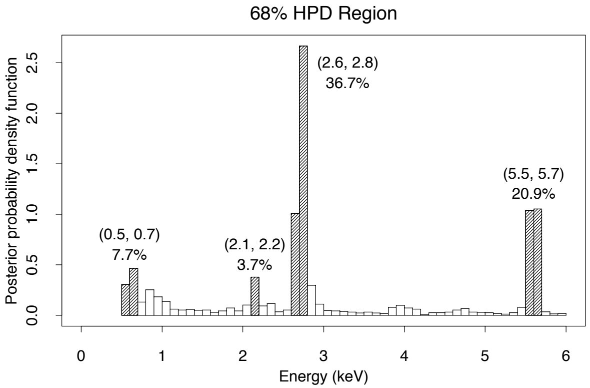
Equal-tailed and HPD intervals for a skewed gamma dist'n:



The difference is more pronounced for more extreme distributions!

Choice of Posterior Intervals (con't)

For a multimodal posterior, HPD may not be an interval! [1]



[1] See Park, van Dyk, and Siemiginowska (2008). Searching for Narrow Emission Lines in X-ray Spectra: Computation and Methods. *ApJ*, **688**, 807–825.

Predictive Distribuitons

The Prior Predictive Distribution: Let y_{rep} be new data.

$$p(y_{\text{rep}}) = \int p(\theta, y_{\text{rep}}) d\theta = \int p_Y(y_{\text{rep}}|\theta) p(\theta) d\theta$$

- Used primarily for model comparison.
- Also called *marginal distribution of data* or *Bayesian evidence*.

The Posterior Predictive Distribution:

$$p(y_{\text{rep}}|y) = \int p(y_{\text{rep}}, \theta|y) d\theta = \int p(y_{\text{rep}}|\theta, y) p(\theta|y) d\theta = \int p(y_{\text{rep}}|\theta) p(\theta|y) d\theta$$

- Used for prediction (and model validation).
- We assume $\tilde{y}$ and y are independent given θ.
- Compare predictive dist'ns in terms of Monte Carlo sample.

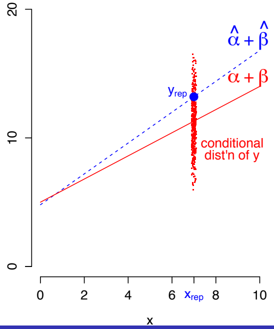
Benefits of Mathematical Foundation

Once we have established $p(y|\theta)$ and $p(\theta)$, everything follows from basic probability theory.

EXAMPLE: Full accounting of uncertainty.

Let $y_i = \alpha + \beta x_i + e_i$, and $e_i \sim \text{NORM}(0, \sigma^2)$ for $i = 1, \ldots, n$.

- New data: $y_{\text{rep}} = \alpha + \beta x_{\text{rep}} + e_{\text{rep}}$
- Prediction: $\hat{y}_{\text{rep}} = \hat{\alpha} + \hat{\beta} x_{\text{rep}}$.
- Two sources of error
  - $\hat{\alpha}$ and $\hat{\beta}$ are only estimates.
  - residuals: $e_{\text{rep}} \sim \text{NORM}(0, \sigma^2)$
- Posterior predictive distribution automatically incorporates both.



Benefits of Mathematical Foundation (con't)

EXAMPLE: The Posterior Odds.

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(y|\theta_1)p(\theta_1)/p(y)}{p(y|\theta_2)p(\theta_2)/p(y)} = \frac{p(y|\theta_1)}{p(y|\theta_2)} \times \frac{p(\theta_1)}{p(\theta_2)}$$

$$= \text{likelihood ratio} \times \text{prior odds}\ .$$

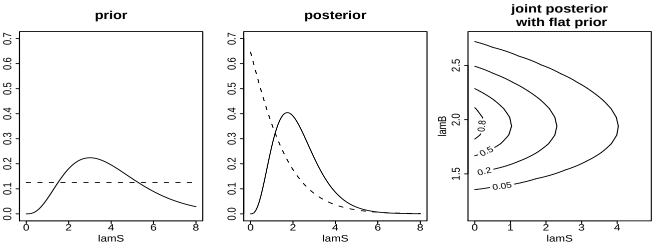
1. Used to compare two parameter values of interest.
2. Geneses of Bayesian methods for model comparison.
3. No new methods required, just standard probability calculations.

Nuissance Parameters

Summarizing the posterior distribution:

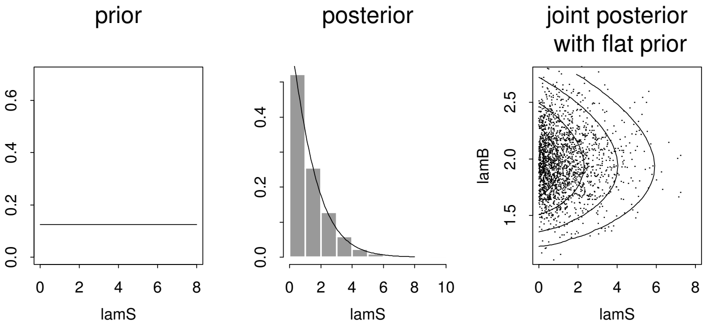
- We can plot the contours of the posterior distribution.
- Plot the marginal distributions of the parameters of interest:

$$p(\lambda_S \mid y, y_B) = \int p(\lambda_S, \lambda_B \mid y, y_B) d\lambda_B$$



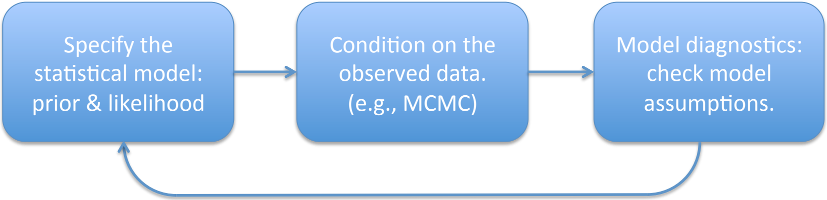
Markov Chain Monte Carlo

Exploring the posterior distribution via Monte Carlo.



Easily generalizes to higher dimensions.

Bayesian Data Analysis: The Big Picture



- Statisticians: Model checking and model improvement.
- Scientists: Model comparison and model selection.

But remember....

All models are wrong, but some are useful.

—George Box

Outline

Bayesian Analysis of Standard Binomial Model

EXAMPLE: Hardness Ratios in High Energy Astrophysics[2]

Let

- $H \sim \textsc{Poisson}(\lambda_H)$ be the observed hard count.
- $S \sim \textsc{Poisson}(\lambda_S)$ be the observed soft count.
- $n = H + S$ be the total count.

If H and S are independent,

$$H|n \sim \textsc{Binomial}\left(n, \pi = \frac{\lambda_H}{\lambda_H + \lambda_S}\right)$$

We will conduct a Bayesian Analysis of this model, treating π as the unknown parameter.

[2] For more on Bayesian analysis of Hardness Ratios see Park et al. (2006). Hardness Ratios with Poisson Errors: Modeling and Computations. *ApJ*, **652**, 610–628.

Details of Binomial Analysis

Likelihood:

$$p_H(h|\pi) = \frac{n!}{h!(n-h)!}\pi^h(1-\pi)^{n-h} \text{ for } h = 0, 1, \ldots, n$$

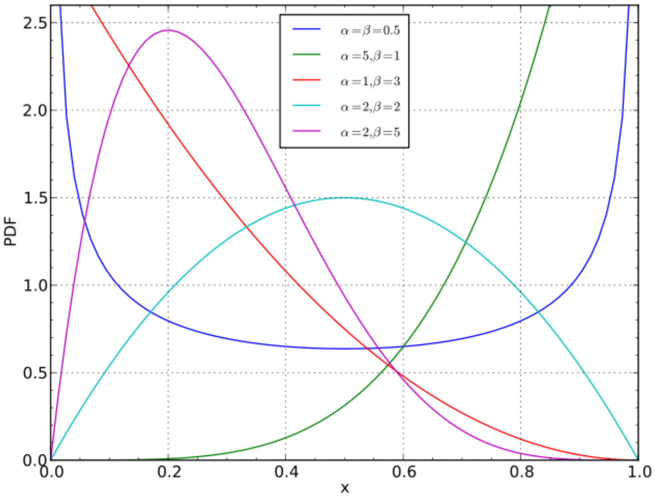
Beta prior distribution:

$$p(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \text{ for } 0 < \pi < 1$$

where α and β are hyper parameters, which define prior dist'n.

The beta family is a flexible class of prior distributions on the unit interval.

Beta Distributions: A Flexible Class of Priors



Beta Dist'n is Conjugate to the Binomial

$$\boxed{\begin{array}{c}\text{If } H|n,\pi \overset{\text{dist}}{\sim} \textsc{Binomial}(n,\pi) \text{ and } \pi \overset{\text{dist}}{\sim} \textsc{Beta}(\alpha,\beta) \\ \text{then } \pi|H,n \overset{\text{dist}}{\sim} \textsc{Beta}(h+\alpha, n-h+\beta).\end{array}}$$

Suppressing the conditioning on n,

$$\begin{aligned} p(\pi|h) &\propto p(h|\pi)\, p(\pi) \\ &= \frac{n!}{h!(n-h)!}\pi^h(1-\pi)^{n-h} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &\propto \pi^{h+\alpha-1}(1-\pi)^{n-h+\beta-1}, \end{aligned}$$

which is proportional to a $\textsc{Beta}(h+\alpha, n-h+\beta)$ density.

Beta Dist'n is Conjugate to the Binomial

$$\boxed{\begin{array}{c}\text{If } H|n,\pi \overset{\text{dist}}{\sim} \textsc{Binomial}(n,\pi) \text{ and } \pi \overset{\text{dist}}{\sim} \textsc{Beta}(\alpha,\beta) \\ \text{then } \pi|H,n \overset{\text{dist}}{\sim} \textsc{Beta}(h+\alpha, n-h+\beta).\end{array}}$$

NOTE:

- The posterior distribution is an "average" of the data/likelihood and the prior distribution.
- We can interpret the hyperparameters α and β as "prior hard and soft counts".
- As n increases, choice of prior matters less.
- Point estimate for π:

$$\mathrm{E}(\pi|h) = \frac{h+\alpha}{n+\alpha+\beta}$$

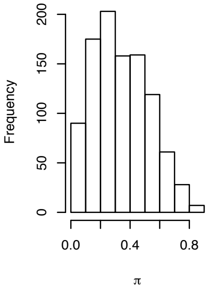
But be cautious of summarizing a dist'n with its mean!

Sample R code

```
# set (flat) prior
> alpha <- 1
> beta <- 1
>
> # set data
> hard <- 1
> soft <- 3
>
> # Monte Carlo sample of posterior
> post.sample.pi <- rbeta(1000, hard + alpha, soft +beta)
>
> estimate <- mean(post.sample.pi)
> error.bar <- sd(post.sample.pi)
> lower <- sort(post.sample.pi)[25]
> upper <-sort(post.sample.pi)[975]
>
> hist(post.sample.pi, xlab =expression(pi), main="")
```

Sample R output

```
> estimate
0.3237472
> error.bar
0.1719679
> lower
0.05146435
> upper
0.6926952
```



Two 95% intervals

- estimate $\pm\, 2\times$ error bars: $(-0.02, 0.66)$
- equil-tail: $(0.05, 0.69)$

Why the difference?

Outline

Parameterization of Hardness Ratio

We have formulated our analysis of Hardness ratios in terms of

$$\pi = \frac{\lambda_H}{\lambda_H + \lambda_S}.$$

Other formulations are more common:

$$\text{simple ratio: } \mathcal{R} = \frac{\lambda_S}{\lambda_H} = \frac{1-\pi}{\pi}$$

$$\text{color: } \mathrm{C} = \log_{10}\left(\frac{\lambda_S}{\lambda_H}\right) = \log_{10}(1-\pi) - \log_{10}(\pi)$$

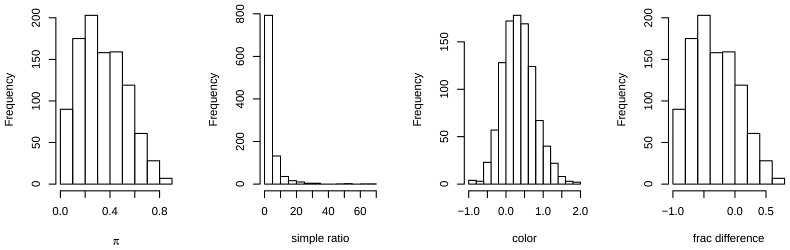
$$\text{fractional difference: } \mathcal{HR} = \frac{\lambda_H - \lambda_S}{\lambda_H + \lambda_S} = 2\pi - 1$$

Transformations of scale and/or parameter are common.

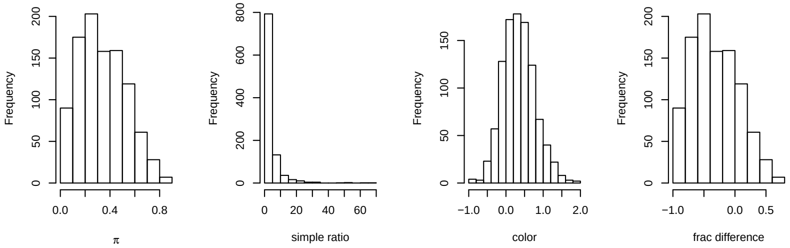
Parameterization of Hardness Ratio

With an MC sample from posterior, transformations are trivial:

```
# Monte Carlo sample of posterior of transformed parameters
> post.sample.ratio <- (1-post.sample.pi)/post.sample.pi
> post.sample.color <- log10(post.sample.ratio)
> post.sample.diff <- 2*post.sample.pi - 1
```



Parameterization of Hardness Ratio



- How will the equal tail intervals compare with that for π?
- How will the HPD intervals compare?
- How will the “estimate $\pm$ 2$\times$ error bar” interval compare?
- What transformation is “best” from a stats perspective?

Outline

Interpreting prior distributions

Using hardness ratios for illustration,

1. POPULATION/FREQUENCY INTERPRETATION: Imagine a population of sources, experiments, or universes from which the current parameter is draw.

   "This source is drawn from a population of sources."

2. STATE OF KNOWLEDGE: A subjective probability dist'n.
3. LACK OF KNOWLEDGE: UNIFORM$(0, 1)$ corresponds to "no prior information". This choice of prior does draw $\mathrm{E}(\pi|h)$ toward $1/2$, but has relatively large prior variance.

We refer to "subjective" and "objective" Bayesian methods

Objective Bayesian Methods

Definition

A reference prior is a prior distribution than can be used as a matter of course under a given likelihood. That is, once the likelihood is specified the reference prior can be automatically applied.

Reference priors might be formulated to

1. minimize the information conveyed by the prior, or
2. optimize other statistical properties of estimators.

For example, we may find the prior that maximizes

$$\mathrm{Var}(\theta|y) \text{ (for all } y \text{ and/or transformation of } \theta??)$$

or yields confidence intervals with correct frequency coverage.

Non-informative Prior Distributions

Definition

A non-informative prior is a prior that aims to play a minimal role in the statistical inference.

Common choice: flat or uniform prior over range of parameter.

EXAMPLE: $h \mid \pi \sim \textsc{Binomial}(n, \pi)$ with $\pi \sim \textsc{Uniform}(0, 1)$.
What does this choice of prior correspond to for:

$$\text{simple ratio: } \mathcal{R} = \frac{\lambda_S}{\lambda_H} = \frac{1-\pi}{\pi}$$

$$\text{color: } C = \log_{10}\left(\frac{\lambda_S}{\lambda_H}\right) = \log_{10}(1-\pi) - \log_{10}(\pi)$$

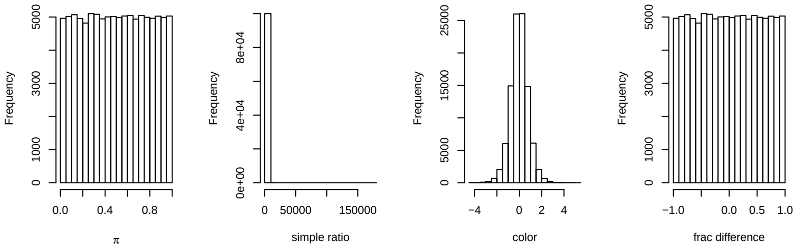
$$\text{fractional difference: } \mathcal{HR} = \frac{\lambda_H - \lambda_S}{\lambda_H + \lambda_S} = 2\pi - 1$$

The Effect of Transformation on the Prior

R-code for an Monte Carlo study:

```
> prior.sample.pi <- runif(100000,0,1)
>
> # Monte Carlo sample of prior of transformed parameters
> prior.sample.ratio <- (1-prior.sample.pi)/prior.sample.pi
> prior.sample.color <- log10(prior.sample.ratio)
> prior.sample.diff <- 2*prior.sample.pi -1
>
> # Histograms
> pdf("hr-2.pdf", width=8, height=3)
> par(mfrow=c(1,4))
> hist(prior.sample.pi, xlab =expression(pi), main="")
> hist(prior.sample.ratio, xlab = "simple ratio", main="")
> hist(prior.sample.color, xlab = "color", main="")
> hist(prior.sample.diff, xlab = "frac difference", main="")
> dev.off()
```

Effect of Transformation on the Prior (cont)



- While the idea of a “flat prior dist’n” seem sensible enough, it is completely determined by the choice of parameter.
- Color is a standard normalizing transformation in stats.[3]
- Why not use flat prior on $\psi =$ color: $p(\psi) \propto 1$ for $-\infty < \psi < \infty$?

[3] But statisticians call $\ln(\pi/(1-\pi))$ the log odds.

Improper Prior Distributions

Definition

An improper prior distribution is a positive-valued function that is not integrable, but that is used formally as a prior distribution.

NOTE:

- Because improper priors are not distributions, we can not rely on probability theory alone.
- However, improper priors generally cause no problem so long as we verify that the resulting posterior distribution is a proper distribution.
- If the posterior distribution is not proper, no sensible conclusions can be drawn.

Example of an Improper Prior Distribution

$$\boxed{\begin{array}{c}\text{If } H|n,\pi \overset{\text{dist}}{\sim} \textsc{Binomial}(n,\pi) \text{ and } \pi \overset{\text{dist}}{\sim} \textsc{Beta}(\alpha,\beta) \\ \text{then } \pi|H,n \overset{\text{dist}}{\sim} \textsc{Beta}(h+\alpha, n-h+\beta).\end{array}}$$

The flat improper prior distribution on color:

$$p(\phi) \propto 1 \text{ for } -\infty < \phi < \infty$$

$[\phi = \mathrm{C} = \log_{10}(\lambda_S/\lambda_H)]$

corresponds to the (improper) distribution on π

$$\pi \sim Beta(\alpha = 0, \beta = 0).$$

The posterior distribution, however, is proper so long as

1. $h \geq 1$ and
2. $n - h \geq 1$.

Example of an Improper Prior Distribution

$$\boxed{\begin{array}{c}\text{If } H|n,\pi \stackrel{\text{dist}}{\sim} \textsc{Binomial}(n,\pi) \text{ and } \pi \stackrel{\text{dist}}{\sim} \textsc{Beta}(\alpha,\beta) \\ \text{then } \pi|H,n \stackrel{\text{dist}}{\sim} \textsc{Beta}(h+\alpha, n-h+\beta).\end{array}}$$

For a general beta prior, the posterior variance of π is:

$$\mathrm{Var}(\pi \mid h, n) = \frac{(h+\alpha)(n-h+\beta)}{(n+\alpha+\beta)^2\,(n+\alpha+\beta+1)}$$

Objective Bayes: Choose α and β to maximize $\mathrm{Var}(\pi \mid h, n)$.

What is the objective choice?

Jeffrey's Invariance Principle

Question: Can we find an objective rule for generating priors that does not depend on the choice of parameterization?

Definition

Jeffery's invariance principle says that any rule for determining a (non-informative) prior distribution should yield the same result if applied to a transformation of the parameter.

NOTE: Any subjective prior distribution should adhere to Jeffery's invariance principle. (At least in principle.)

Jeffrey's Prior Distribution

In likelihood-based statistics, the *Expected Fisher Information* is

$$J(\theta) = -\mathrm{E}\left[\frac{\mathrm{d}^2 \log p(y|\theta)}{\mathrm{d}^2\theta} \;\middle|\; \theta\right]$$

Definition

The Jeffery's prior distribution is

$$p(\theta) \propto \sqrt{J(\theta)}$$

or in higher dimensions,

$$p(\theta) \propto \sqrt{|J(\theta)|}.$$

Example of Jeffrey's Prior

Example: For the binomial model,

$$\log(p_H(h|\pi)) = h\log(\pi) + (n-h)\log(1-\pi) + \text{ constant }.$$

and the expected Fisher information is

$$-\mathrm{E}\left[-\frac{h}{\pi^2} - \frac{n-h}{(1-\pi)^2} \;\Big|\; \pi\right] = \frac{n}{\pi(1-\pi)}.$$

So the Jeffrey's Prior is

$$p(\pi) \propto \sqrt{J(\pi)} \propto \pi^{-1/2}(1-\pi)^{-1/2} = \textsc{Beta}(\alpha = 1/2, \beta = 1/2).$$

This prior is invariant, but is it non-informative??

Are Uniform Priors Non-Informative?

Suppose $$\pi \sim \textsc{Uniform}(0, 1).$$

Question: What is the probability that π is in interval $(\epsilon, 1-\epsilon)$?

Now suppose we have a vector of K parameters with

$$\pi_k \overset{\text{indep}}{\sim} \textsc{Uniform}(0, 1) \text{ for } k = 1, \ldots, K.$$

Question: What is probability that all π_k are in interval $(\epsilon, 1-\epsilon)$?

$$(1-2\epsilon)^k \to 0 \text{ as } k \to \infty$$

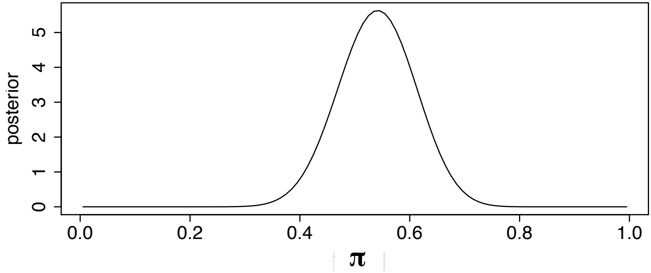
Where is all of the probability??

Prior/Likelihood Mismatch

$$\boxed{\text{If } H|n,\pi \overset{\text{dist}}{\sim} \text{BINOMIAL}(n,\pi) \text{ and } \pi \overset{\text{dist}}{\sim} \text{BETA}(\alpha,\beta) \text{ then } \pi|H,n \overset{\text{dist}}{\sim} \text{BETA}(h+\alpha, n-h+\beta).}$$

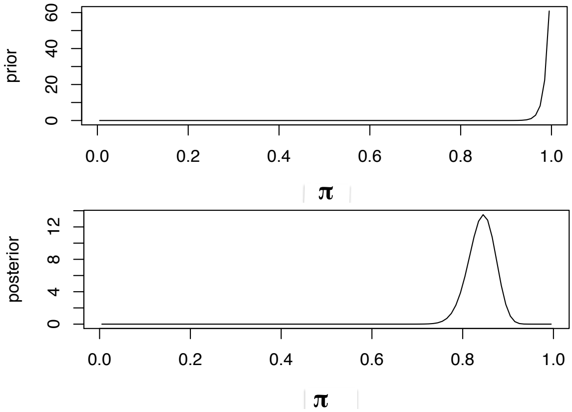
Consider larger dataset: $n = 48$ counts w/ $h = 26$ hard counts.

Prior I: $\pi \sim \text{BETA}(1,1)$ yields:



Prior/Likelihood Mismatch (con't)

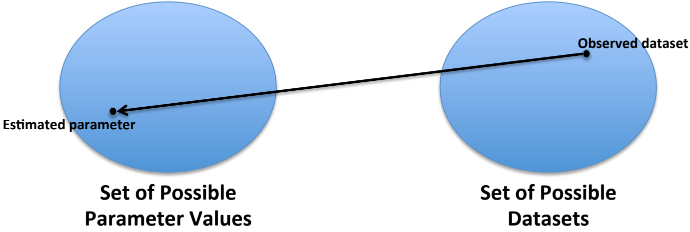
Prior II: $\pi \sim \text{BETA}(1000, 1)$:



In this case $\text{Var}(\pi|h) > \text{Var}(\pi)$.

Outline

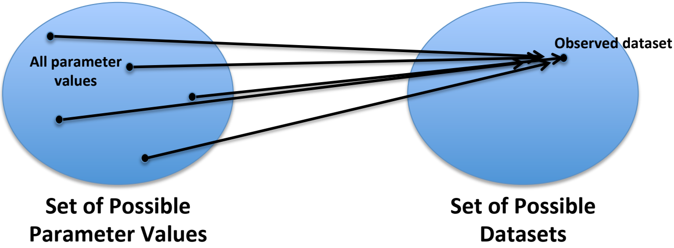
The Goal of Parameter Estimation



Given the observed dataset:

- Find the most likely or most probably value of parameter.
- Find an estimate that is likely to be near the “true” value of the parameter.

Likelihood-based Inference



Draws the arrows in the *wrong* direction:

- For each value of the parameter how likely would the observed data be?

Reversing the conditioning in a probabilistic statement can be highly misleading!

Reversing the Conditioning

Question:

- Can we conclude that there is evidence for new physics?
- Does Pr(data | new physics) large imply

  Pr(new physics | data) is large??

Order of conditioning matters!

Consider $\Pr(A \mid B)$ and $\Pr(B \mid A)$ with

A:

B:

Reversing the Conditioning

Question:

- Can we conclude that there is evidence for new physics?
- Does Pr(data | new physics) large imply
  Pr(new physics | data) is large??

Order of conditioning matters!

Consider Pr(A | B) and Pr(B | A) with

A: A person is a woman.

B: A person is pregnant.

The Values may be very different!

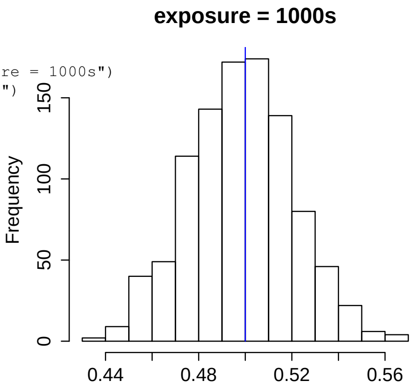
Justification for Likelihood-based Inference

Asymptotic frequency properties:

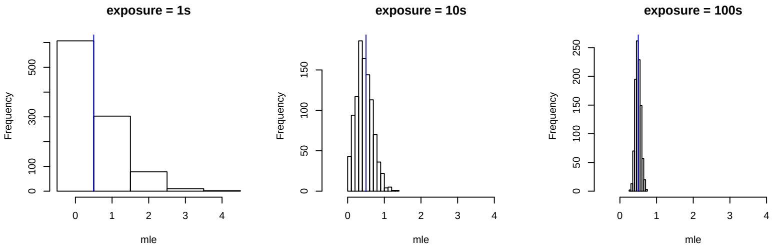
- If you consider the data to be a random sample of possible data sets, the MLE, $\hat{\theta}_{\rm mle}$ is also random.
- Because it is a random quantity, we can compute the distribution, mean, and variance of $\hat{\theta}_{\rm mle}$.
- If the size of the data is *LARGE* (asymptotic!), then
  1. Mean of $\hat{\theta}_{\rm mle}$ is near its true value (MLE is asy. unbiased).
  2. Variance of $\hat{\theta}_{\rm mle}$ decreases as sample size increases.
  3. The distribution of $\hat{\theta}_{\rm mle}$ is approximately Gaussian (MLE is asymptotically Gaussian).

Example of Asymptotic Behavior of MLE

```
> # Number of replicate data sets
> N <- 1000
> exposure<-1000
>
> # Generate replicate data sets and compute mle's
> data <- rpois(N, 0.5*exposure)
> mle <- data / exposure
>
> pdf("asy-1.pdf", width=4, height=4)
> par(mex=0.7)
> hist(mle, xlab = "mle", main="exposure = 1000s")
> lines(rep(0.5,2),c(0,10^6), col="blue")
> dev.off()
```



Changing the Exposure



- The MLE works great for large samples.
- But it has no direct justification in small sample settings.
- Frequency properties must be derived case-by-case.

What about Bayesian Methods?

Bayesian methods have the same asymptotic properties as likelihood-based methods
(as long as prior has some probability around the true value).

In addition Bayesian methods

1. have probabilistic justification in small samples (w/out asymptotics),
2. can be justified in terms of small sample frequency properties on a case-by-case basis,
3. are much easier to interpret using probability statements,
4. naturally allow for multiple sources of information.

Choosing the Prior Distribution

Solance: *Any reasonable prior distribution* results in exactly the same asymptotic frequency properties as likelihood methods.

Worry: Only if you want to do better than likelihood-based methods in small samples.

Diligence: Nonetheless in practice much effort is put into selecting priors that help us best achieve our objectives.

Advantage: The choice of prior is an additional degree of freedom in methodological development.

Choice of prior can even improve frequency properties!

Frequency Properties of Bayesian Methods

EXAMPLE: Suppose $H \sim \text{BINOMIAL}(n = 10, \pi)$.

Consider four estimates of π:

i) $\hat{\pi}_1$, the maximum likelihood estimator of π;
ii) $\hat{\pi}_2 = \mathrm{E}(\pi|Y)$, where π has prior distribution $\pi \sim \mathrm{Beta}(1, 1)$
iii) $\hat{\pi}_3 = \mathrm{E}(\pi|Y)$, where π has prior distribution $\pi \sim \mathrm{Beta}(1, 4)$
iv) $\hat{\pi}_4 = \mathrm{E}(\pi|Y)$, where π has prior distribution $\pi \sim \mathrm{Beta}(4, 1)$

and four 95% interval estimators of π,

$$\hat{\pi}_i \pm 1.96\sqrt{\frac{1}{n}\hat{\pi}_i(1 - \hat{\pi}_i)} \quad \text{for} \quad i = 1, \ldots, 4.$$

Frequency Properties of Estimators and Intervals

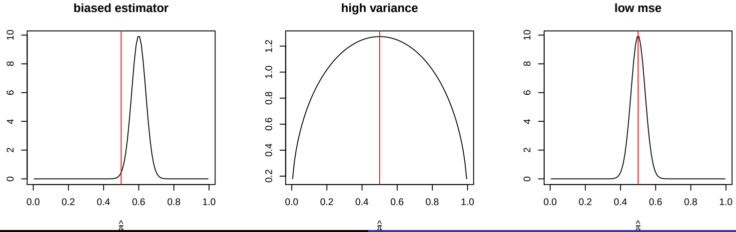
Remember: If the data is a random sample of all possible data, the estimator $\hat{\pi}_i$ is also random. It has a distribution, mean, and variance.

We can evaluate the $\hat{\pi}_i$ as an estimator of π in terms of its

bias: $\mathrm{E}(\hat{\pi}_i \mid \pi) - \pi$ (Is bias bad??)

variance: $\mathrm{E}\left[\left(\hat{\pi}_i - \mathrm{E}(\hat{\pi}_i \mid \pi)\right)^2 \mid \pi\right]$

mean square error: $\mathrm{E}\left[(\hat{\pi}_i - \pi)^2 \mid \pi\right] = \text{bias}^2 + \text{variance}$



You will investigate the frequency properties of Bayesian estimates of a Binomial probability in the first lab.

Subjective vs. Objective Analysis

All *statistical analyses are subjective.* Choices of data, parametric forms, statistical/scientifc models, "what to model".

But Bayesian methods have one more subjective component, the quantification of prior knowledge in through a distribution.

And prior distributions need't be used in subjective manner.

Everything follows from basic probability theory once we have established $p(y|\theta)$ and $p(\theta)$, Compare with likelihood theory.

Asymptotic results and counter intuitive definitions (e.g., for a CI or a p-value) *are not required*.