Introduction to Bayesian Inference: Supplemental Topics

Tom Loredo
Dept. of Astronomy, Cornell University
http://www.astro.cornell.edu/staff/loredo/bayes/

CASt Summer School — 5 June 2014

Supplemental Topics

1. Parametric bootstrapping vs. posterior sampling
2. Binary classification and case diagrams
3. Estimation and model comparison for binary outcomes
4. Student's t distribution via marginalization
5. Two more solutions of the on/off problem

Supplemental Topics

1. **Parametric bootstrapping vs. posterior sampling**
2. Binary classification and case diagrams
3. Estimation and model comparison for binary outcomes
4. Student's t distribution via marginalization
5. Two more solutions of the on/off problem

Likelihood-Based Parametric Bootstrapping

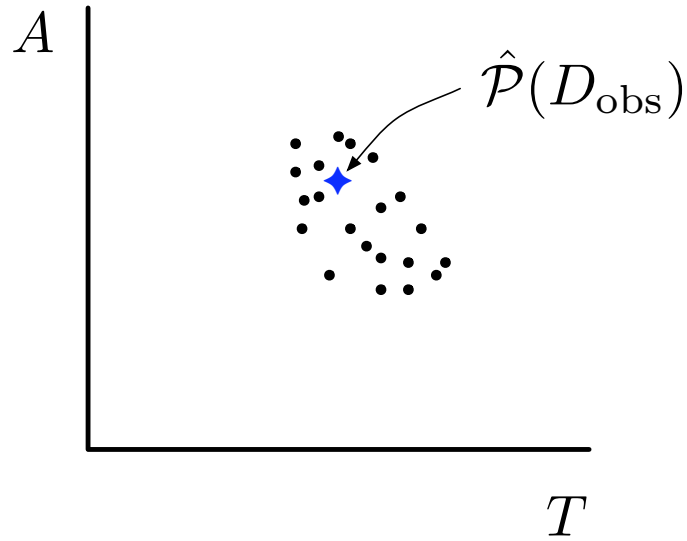
Likelihood $\mathcal{L}(\theta) \equiv p(D_{\text{obs}}|\theta)$.
Log-likelihood $L(\theta) = \ln \mathcal{L}(\theta)$.

For the Gaussian example,

$$\begin{aligned}
\mathcal{L}(\mu) &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&\propto \prod_i \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
L(\mu) &= -\frac{1}{2}\sum_i \frac{(x_i-\mu)^2}{\sigma^2} + \text{Const} \\
&= -\frac{\chi^2(\mu)}{2} + \text{Const}
\end{aligned}$$

Incorrect Parametric Bootstrapping

$$\mathcal{P} = (A, T)$$

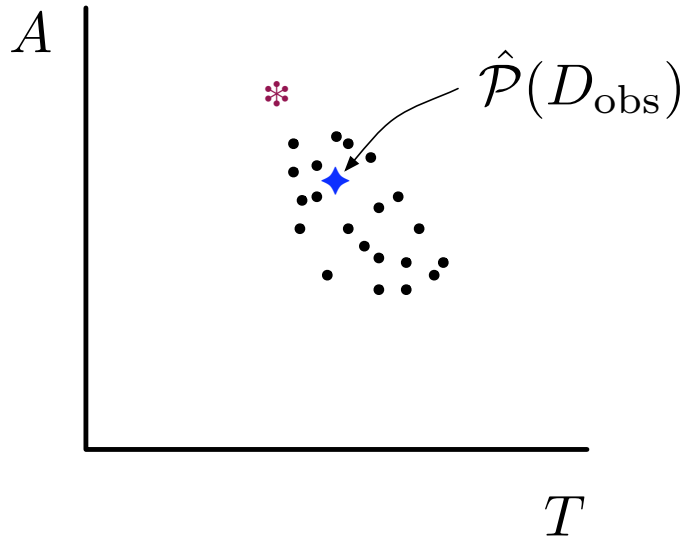


Histograms/contours of best-fit estimates from $D \sim p(D|\hat{\theta}(D_{\text{obs}}))$ provide *poor* confidence regions—no better (possibly worse) than using a least-squares/χ^2 covariance matrix.

What's wrong with the population of $\hat{\theta}$ points for this purpose?

Incorrect Parametric Bootstrapping

$$\mathcal{P} = (A, T)$$



Histograms/contours of best-fit estimates from $D \sim p(D|\hat{\theta}(D_{\text{obs}}))$ provide *poor* confidence regions—no better (possibly worse) than using a least-squares/χ^2 covariance matrix.

What's wrong with the population of $\hat{\theta}$ points for this purpose?

The estimates are skewed down and to the right, indicating the truth must be up and to the left. *Do not mistake variability of the estimator with the uncertainty of the estimate!*

Key idea: Use likelihood *ratios* to define confidence regions.
I.e., use L or χ^2 *differences* to define regions.

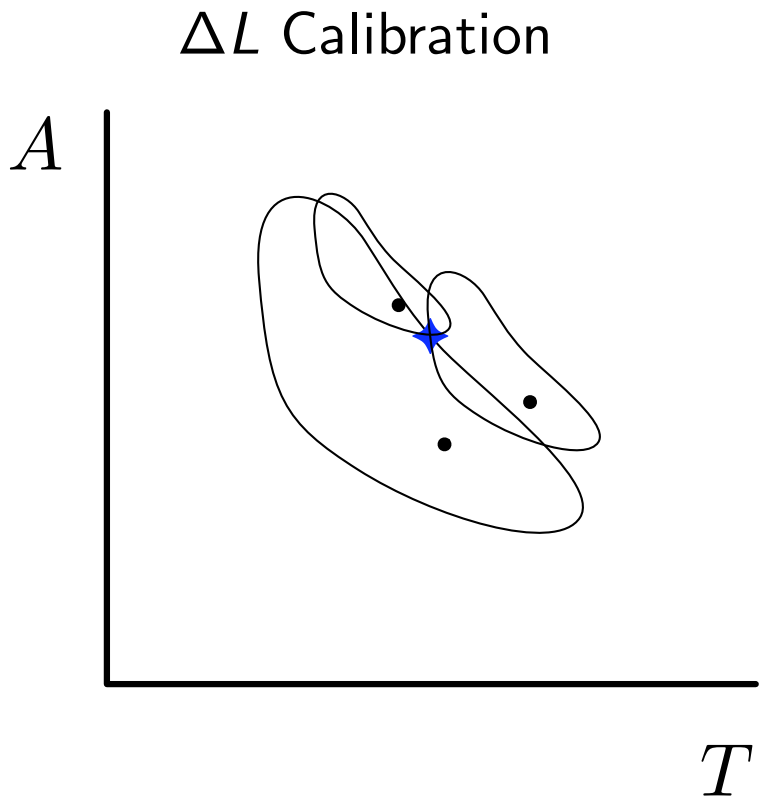
Estimate parameter values via *maximum likelihood* ($\min \chi^2$) $\rightarrow L_{\max}$.
Pick a constant ΔL. Then

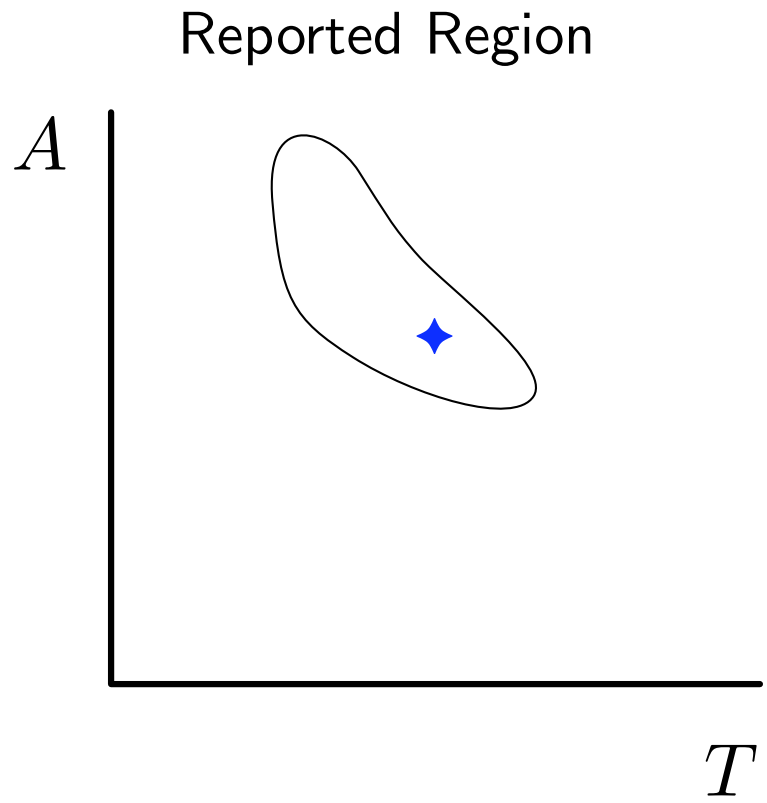
$$\Delta(D) = \{\theta : L(\theta) > L_{\max} - \Delta L\}$$

Coverage calculation:

1. Fix $\theta_0 = \hat{\theta}(D_{\text{obs}})$ (plug-in approx'n)
2. Simulate a dataset from $p(D|\theta_0) \rightarrow L_D(\theta)$
3. Find maximum likelihood estimate $\hat{\theta}(D)$
4. Calculate $\Delta L = L_D(\hat{\theta}_D) - L_D(\theta_0)$
5. Goto (2) for N total iterations
6. Histogram the ΔL values to find coverage vs. ΔL (fraction of sim'ns with smaller ΔL)

Report $\Delta(D_{\text{obs}})$ with ΔL chosen for desired approximate CL.





The CL is approximate due to:

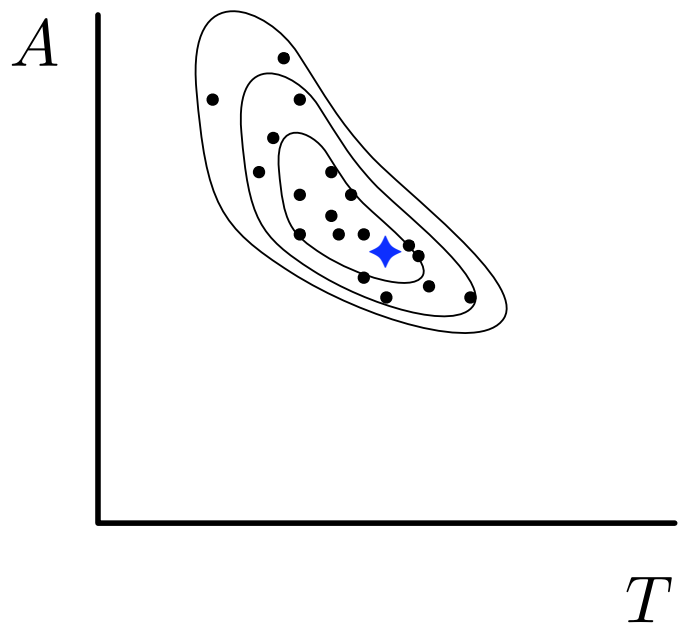
- Monte Carlo error in calibrating ΔL
- The plug-in approximation

Credible Region Via Posterior Sampling

Monte Carlo algorithm for finding credible regions:

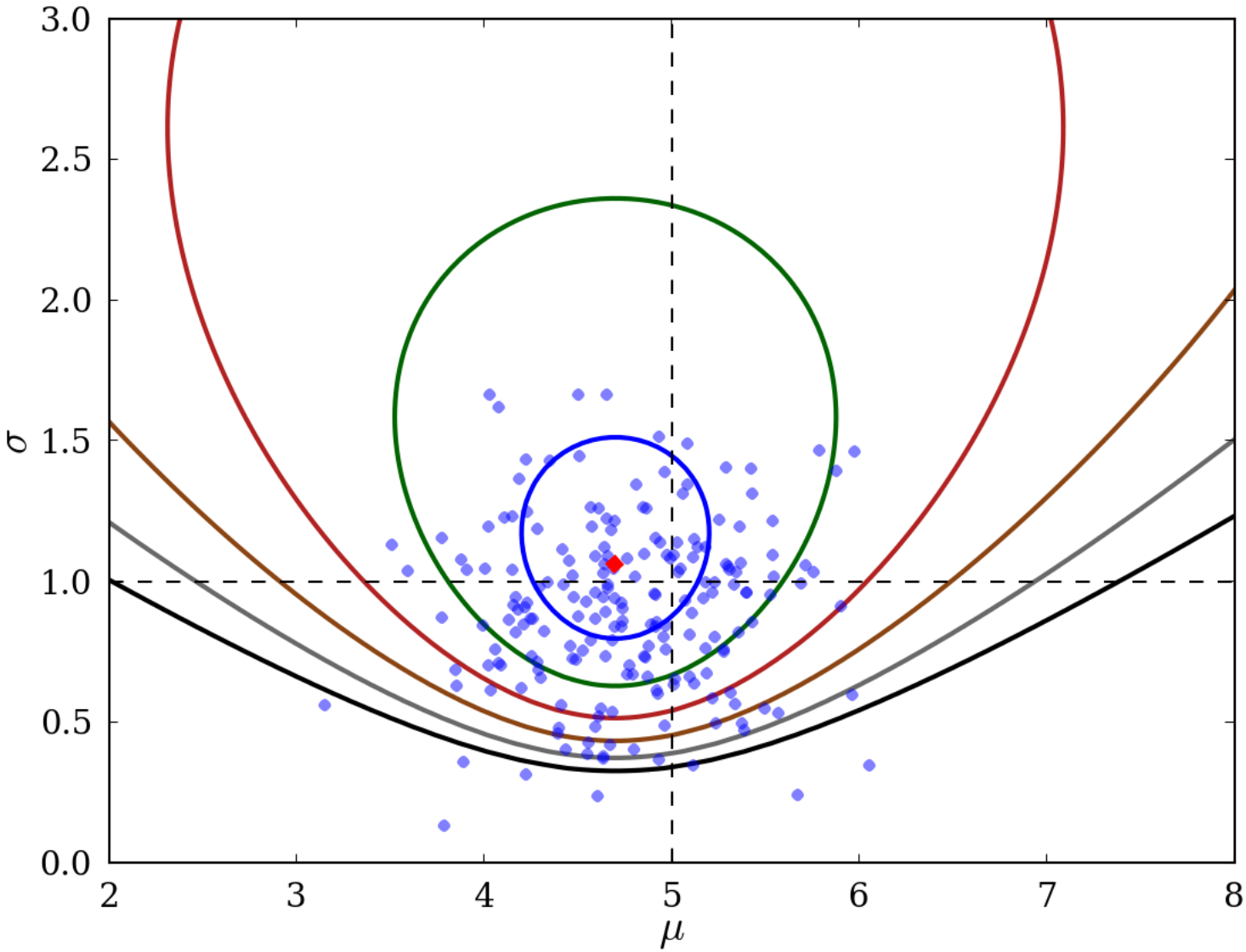
1. Create a RNG that can sample θ from $p(\theta|D_{\text{obs}})$
2. Draw N samples; record θ_i and $q_i = \pi(\theta_i)\mathcal{L}(\mu_i)$
3. Sort the samples by the q_i values
4. An HPD region of probability P is the θ region spanned by the $100P\%$ of samples with highest q_i

Note that no dataset other than D_{obs} is ever considered.
P is a property of the *particular interval* reported.



This complication is the rule rather than the exception!

Simple example: Estimate the mean *and standard deviation* of a normal distribution ($\mu = 5$, $\sigma = 1$, $N = 5$; 200 samples):



Supplemental Topics

1. Parametric bootstrapping vs. posterior sampling
2. **Binary classification and case diagrams**
3. Estimation and model comparison for binary outcomes
4. Student's t distribution via marginalization
5. Two more solutions of the on/off problem

Visualizing Bayesian Inference

Simplest case: Binary classification

- 2 hypotheses: $\{H, C\}$
- 2 possible data values: $\{-, +\}$

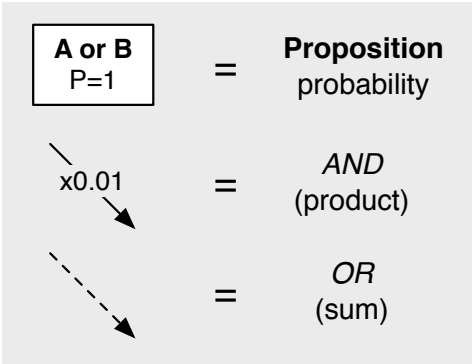
Concrete example: You test positive $(+)$ for a medical condition. Do you have the condition (C) or not $(H$, "healthy"$)$?

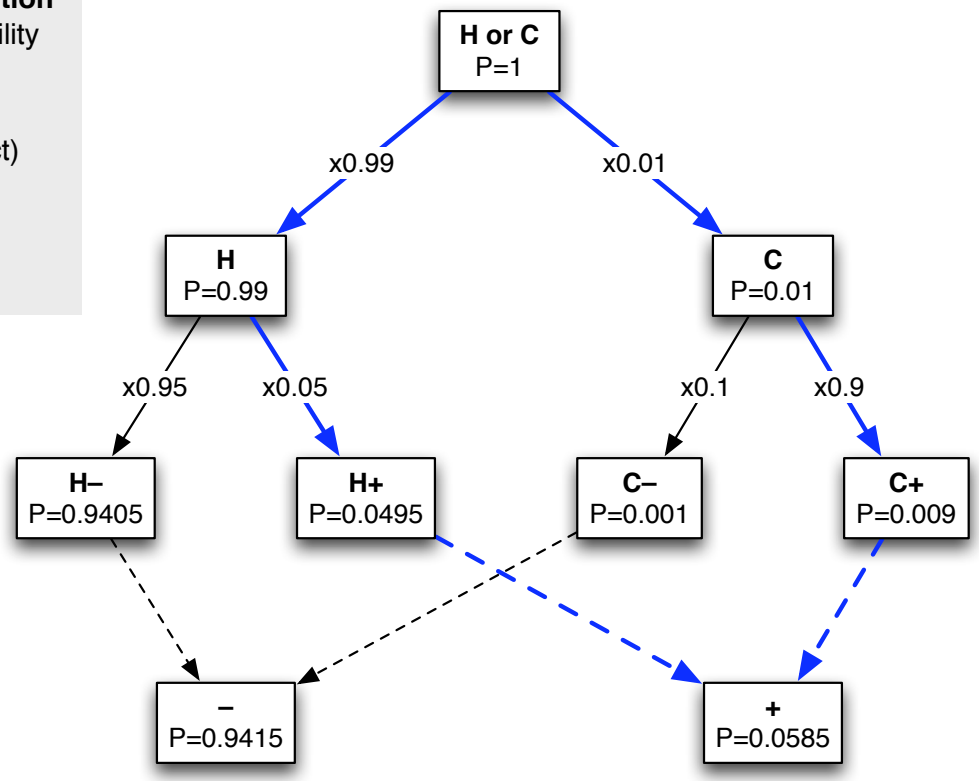
- Prior: Prevalence of the condition in your population is 1%
- Likelihood:
  - Test is 90% accurate if you have the condition: $P(+|C, I) = 0.9$ ("sensitivity")
  - Test is 95% accurate if you are healthy: $P(-|H, I) = 0.95$ ("specificity")

Numbers roughly correspond to breast cancer in asymptomatic women aged 40–50, and mammography screening

[Gigerenzer, *Calculated Risks* (2002)]

Probability "Tree"





$P(H_1 \vee H_2|I)$

$P(H_i|I)$

$P(H_i, D|I) = P(H_i|I)P(D|H_i, I)$

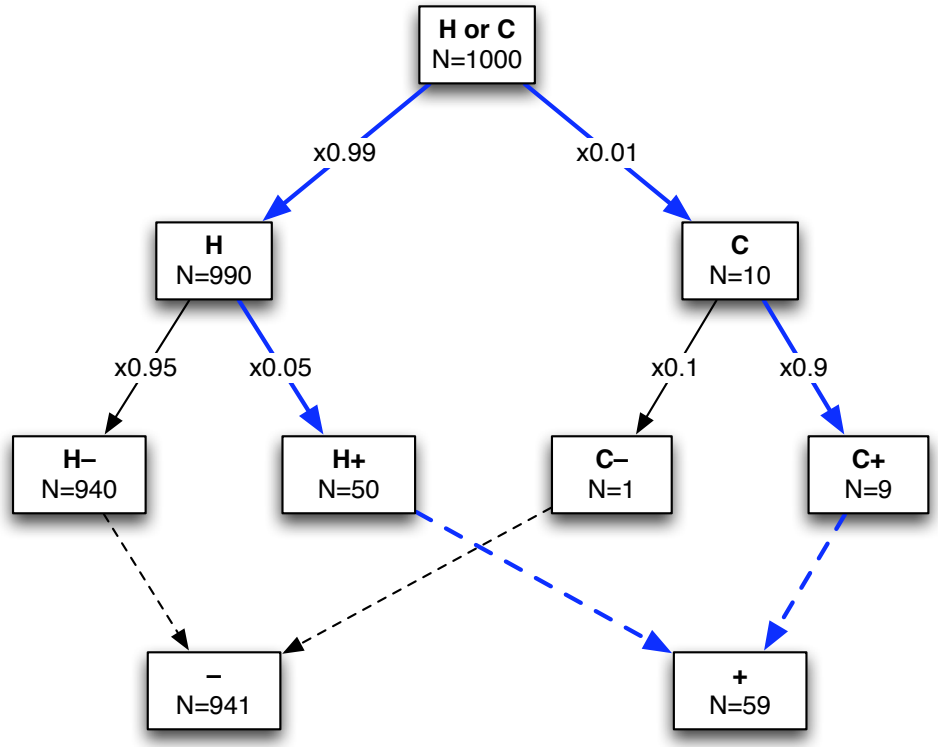
$P(D|I) = \sum_i P(H_i, D|I)$

$$P(C|+, I) = \frac{0.009}{0.0585} \approx 0.15$$

*Not really a tree; really a graph or part of a lattice

Count "Tree"

Integers are easier than reals!
Create a large ensemble of cases so ratios of counts approximate the probabilities.



$$P(C|+,I) = \frac{9}{59} \approx 0.15$$

Of the 59 cases with positive test results, only 9 have the condition. The prevalence is so low that when there is a positive result, it's more likely to have been a mistake than accurate.

Supplemental Topics

1. Parametric bootstrapping vs. posterior sampling
2. Binary classification and case diagrams
3. **Estimation and model comparison for binary outcomes**
4. Student's t distribution via marginalization
5. Two more solutions of the on/off problem

Binary Outcomes: Parameter Estimation

M = Existence of two outcomes, S and F; for each case or trial, the probability for S is α; for F it is $(1 - \alpha)$

H_i = Statements about α, the probability for success on the next trial $\rightarrow$ seek $p(\alpha|D, M)$

D = Sequence of results from N observed trials:

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Likelihood:

$$\begin{aligned} p(D|\alpha, M) &= p(\text{failure}|\alpha, M) \times p(\text{failure}|\alpha, M) \times \cdots \\ &= \alpha^n (1 - \alpha)^{N-n} \\ &= \mathcal{L}(\alpha) \end{aligned}$$

Prior

Starting with no information about α beyond its definition, use as an "uninformative" prior $p(\alpha|M) = 1$. Justifications:

- Intuition: Don't prefer any α interval to any other of same size
- Bayes's justification: "Ignorance" means that before doing the N trials, we have no preference for how many will be successes:

$$P(n\,\text{success}|M) = \frac{1}{N+1} \qquad \rightarrow \qquad p(\alpha|M) = 1$$

Consider this a *convention*—an assumption added to M to make the problem well posed.

Prior Predictive

$$\begin{aligned} p(D|M) &= \int d\alpha\, \alpha^n (1-\alpha)^{N-n} \\ &= B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!} \end{aligned}$$

A *Beta integral*, $B(a,b) \equiv \int dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Posterior

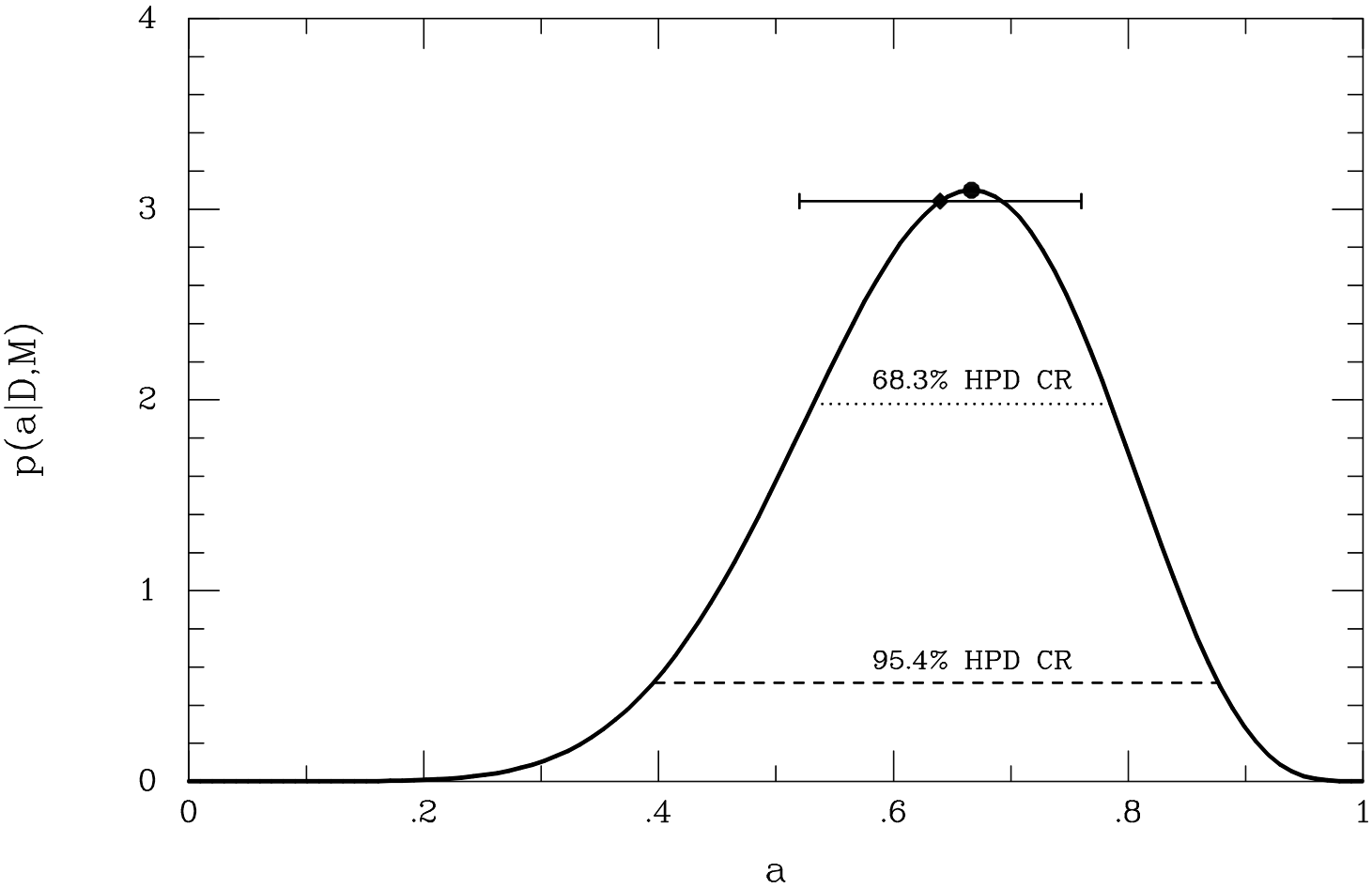
$$p(\alpha|D,M) \quad = \quad \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

A *Beta distribution*. Summaries:

- Best-fit: $\hat{\alpha} = \frac{n}{N} = 2/3$; $\langle\alpha\rangle = \frac{n+1}{N+2} \approx 0.64$
- Uncertainty: $\sigma_\alpha = \sqrt{\frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}} \approx 0.12$
  Find credible regions numerically, or with incomplete beta function

Note that the posterior depends on the data only through n, not the N binary numbers describing the sequence.

n is a (minimal) *sufficient statistic*.


p(a|D,M)
a
68.3% HPD CR
95.4% HPD CR
0
1
2
3
4
0
.2
.4
.6
.8
1

Binary Outcomes: Model Comparison

Equal Probabilities?

M_1: $\alpha = 1/2$

M_2: $\alpha \in [0, 1]$ with flat prior.

Maximum Likelihoods

$$M_1: \qquad p(D|M_1) = \frac{1}{2^N} = 2.44 \times 10^{-4}$$

$$M_2: \quad \mathcal{L}(\hat{\alpha}) \;=\; \left(\frac{2}{3}\right)^n \left(\frac{1}{3}\right)^{N-n} = 4.82 \times 10^{-4}$$

$$\frac{p(D|M_1)}{p(D|\hat{\alpha}, M_2)} = 0.51$$

Maximum likelihoods favor M_2 (failures more probable).

Bayes Factor (ratio of model likelihoods)

$$p(D|M_1) = \frac{1}{2^N}; \qquad \text{and} \qquad p(D|M_2) = \frac{n!(N-n)!}{(N+1)!}$$

$$\rightarrow B_{12} \equiv \frac{p(D|M_1)}{p(D|M_2)} = \frac{(N+1)!}{n!(N-n)!2^N}$$
$$= 1.57$$

Bayes factor (odds) favors M_1 (equiprobable).

Note that for $n = 6$, $B_{12} = 2.93$; for this small amount of data, we can never be very sure results are equiprobable.

If $n = 0$, $B_{12} \approx 1/315$; if $n = 2$, $B_{12} \approx 1/4.8$; for extreme data, 12 flips *can* be enough to lead us to strongly suspect outcomes have different probabilities.

(Frequentist significance tests can reject null for any sample size.)

Binary Outcomes: Binomial Distribution

Suppose $D = n$ (number of heads in N trials), rather than the actual sequence. What is $p(\alpha|n, M)$?

Likelihood

Let $\mathcal{S}$ = a sequence of flips with n heads.

$$\begin{aligned} p(n|\alpha, M) &= \sum_{\mathcal{S}} p(S|\alpha, M)\, p(n|\mathcal{S}, \alpha, M) \\ &= \alpha^n (1-\alpha)^{N-n} C_{n,N} \end{aligned}$$

$p(S|\alpha, M) \to \alpha^n (1-\alpha)^{N-n}$

$p(n|\mathcal{S}, \alpha, M) \to [\![\# \text{ successes} = n]\!]$

$C_{n,N} = \#$ of sequences of length N with n heads.

$$\to p(n|\alpha, M) = \frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

The *binomial distribution* for n given α, N.

Posterior

$$p(\alpha|n,M) = \frac{\frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}}{p(n|M)}$$

$$\begin{aligned} p(n|M) &= \frac{N!}{n!(N-n)!}\int d\alpha\, \alpha^n(1-\alpha)^{N-n} \\ &= \frac{1}{N+1} \end{aligned}$$

$$\rightarrow p(\alpha|n,M) = \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Same result as when data specified the actual sequence.

Another Variation: Negative Binomial

Suppose $D = N$, the number of trials it took to obtain a predifined number of successes, $n = 8$. What is $p(\alpha|N, M)$?

Likelihood

$p(N|\alpha, M)$ is probability for $n-1$ successes in $N-1$ trials, times probability that the final trial is a success:

$$\begin{aligned} p(N|\alpha, M) &= \frac{(N-1)!}{(n-1)!(N-n)!}\alpha^{n-1}(1-\alpha)^{N-n}\alpha \\ &= \frac{(N-1)!}{(n-1)!(N-n)!}\alpha^{n}(1-\alpha)^{N-n} \end{aligned}$$

The *negative binomial distribution* for N given α, n.

Posterior

$$p(\alpha|D,M) = C'_{n,N}\frac{\alpha^n(1-\alpha)^{N-n}}{p(D|M)}$$

$$p(D|M) \quad = \quad C'_{n,N}\int d\alpha\ \alpha^n(1-\alpha)^{N-n}$$

$$\rightarrow p(\alpha|D,M) = \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Same result as other cases.

Final Variation: "Meteorological Stopping"

Suppose $D = (N, n)$, the number of samples and number of successes in an observing run whose total number was determined by the weather at the telescope. What is $p(\alpha|D, M')$?

(M' adds info about weather to M.)

Likelihood

$p(D|\alpha, M')$ is the binomial distribution times the probability that the weather allowed N samples, $W(N)$:

$$p(D|\alpha, M') \quad = \quad W(N)\frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Let $C_{n,N} = W(N)\binom{N}{n}$. We get the *same result* as before!

Likelihood Principle

To define $\mathcal{L}(H_i) = p(D_{\text{obs}}|H_i, I)$, we must contemplate what other data we might have obtained. But the "real" sample space may be determined by many complicated, seemingly irrelevant factors; it may not be well-specified at all. Should this concern us?

Likelihood principle: The result of inferences depends only on how $p(D_{\text{obs}}|H_i, I)$ varies w.r.t. hypotheses. We can ignore aspects of the observing/sampling procedure that do not affect this dependence.

This happens because no sums of probabilities for hypothetical data appear in Bayesian results; Bayesian calculations *condition on* D_{obs}.

This is a sensible property that frequentist methods do not share. Frequentist probabilities are "long run" rates of performance, and depend on details of the sample space that are irrelevant in a Bayesian calculation.

Goodness-of-fit Violates the Likelihood Principle

Theory (H_0)

The number of "A" stars in a cluster should be 0.1 of the total.

Observations

5 A stars found out of 96 total stars observed.

Theorist's analysis

Calculate χ^2 using $\bar{n}_A = 9.6$ and $\bar{n}_X = 86.4$.

Significance level is $p(> \chi^2 | H_0) = 0.12$ (or 0.07 using more rigorous binomial tail area). Theory is accepted.

Observer's analysis

Actual observing plan was to keep observing until 5 A stars seen!

"Random" quantity is $N_{\rm tot}$, not n_A; it should follow the negative binomial dist'n. Expect $N_{\rm tot} = 50 \pm 21$.

$p(> \chi^2 | H_0) = 0.03$. Theory is rejected.

Telescope technician's analysis

A storm was coming in, so the observations would have ended whether 5 A stars had been seen or not. The proper ensemble should take into account $p(\text{storm})\ldots$

Bayesian analysis

The Bayes factor is the same for binomial or negative binomial likelihoods, and slightly favors H_0. Include $p(\text{storm})$ if you want—it will drop out!

Probability & frequency

Frequencies are relevant when modeling repeated trials, or repeated sampling from a population or ensemble.

Frequencies are observables

- When available, can be used to *infer* probabilities for next trial
- When unavailable, can be *predicted*

Bayesian/Frequentist relationships

- Relationships between probability and frequency
- Long-run performance of Bayesian procedures

Probability & frequency in IID settings

Frequency from probability

Bernoulli's law of large numbers: In repeated i.i.d. trials, given $P(\text{success}|\ldots) = \alpha$, predict

$$\frac{n_{\text{success}}}{N_{\text{total}}} \to \alpha \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be interpreted as a frequency distribution.

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" $\to$ First use of Bayes's theorem:

Probability for success in next trial of i.i.d. sequence:

$$\mathrm{E}(\alpha) \to \frac{n_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be estimated from a frequency distribution.

Supplemental Topics

1. Parametric bootstrapping vs. posterior sampling
2. Binary classification and case diagrams
3. Estimation and model comparison for binary outcomes
4. **Student's t distribution via marginalization**
5. Two more solutions of the on/off problem

Estimating a Normal Mean: Unknown σ

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is *unknown*

Parameter space: (μ, σ); seek $p(\mu|D, M)$

Likelihood

$$\begin{aligned} p(D|\mu, \sigma, M) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\ &\propto \frac{1}{\sigma^N} e^{-Q/2\sigma^2} \end{aligned}$$

$$\text{where} \quad Q = N\left[r^2 + (\mu - \overline{d})^2\right]$$

Uninformative Priors

Assume priors for μ and σ are independent.

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant.

Scale invariance $\Rightarrow p(\sigma) \propto 1/\sigma$ (flat in $\log \sigma$).

Joint Posterior for μ, σ

$$p(\mu, \sigma | D, M) \propto \frac{1}{\sigma^{N+1}} e^{-Q(\mu)/2\sigma^2}$$

Marginal Posterior

$$p(\mu|D,M) \propto \int d\sigma \, \frac{1}{\sigma^{N+1}} e^{-Q/2\sigma^2}$$

Let $\tau = \frac{Q}{2\sigma^2}$ so $\sigma = \sqrt{\frac{Q}{2\tau}}$ and $|d\sigma| = \tau^{-3/2}\sqrt{\frac{Q}{2}}\, d\tau$

$$\begin{aligned} \Rightarrow p(\mu|D,M) &\propto 2^{N/2} Q^{-N/2} \int d\tau \, \tau^{\frac{N}{2}-1} e^{-\tau} \\ &\propto Q^{-N/2} \end{aligned}$$

Write $Q = Nr^2\left[1 + \left(\frac{\mu - \overline{d}}{r}\right)^2\right]$ and normalize:

$$p(\mu|D,M) = \frac{\left(\frac{N}{2}-1\right)!}{\left(\frac{N}{2}-\frac{3}{2}\right)!\sqrt{\pi}}\frac{1}{r}\left[1 + \frac{1}{N}\left(\frac{\mu-\overline{d}}{r/\sqrt{N}}\right)^2\right]^{-N/2}$$

"Student's t distribution," with $t = \frac{(\mu-\overline{d})}{r/\sqrt{N}}$

A "bell curve," but with power-law tails

Large N:

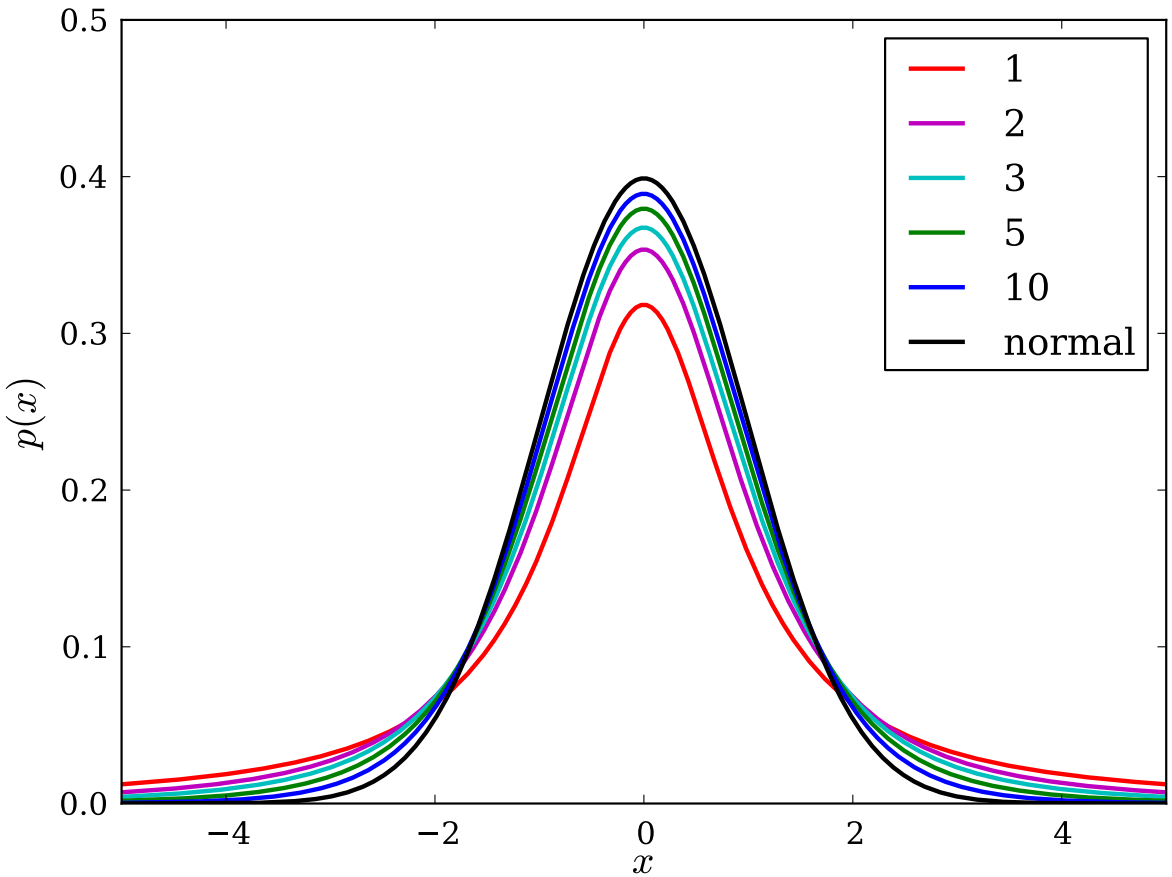
$$p(\mu|D,M) \sim e^{-N(\mu-\overline{d})^2/2r^2}$$

This is the rigorous way to "adjust σ so $\chi^2/\mathrm{dof} = 1$."

It doesn't just plug in a best σ; it slightly broadens posterior to account for σ uncertainty.

Student t examples:

- $p(x) \propto \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$
- Location $= 0$, scale $= 1$
- Degrees of freedom $= \{1, 2, 3, 5, 10, \infty\}$



Supplemental Topics

1. Parametric bootstrapping vs. posterior sampling
2. Binary classification and case diagrams
3. Estimation and model comparison for binary outcomes
4. Student's t distribution via marginalization
5. **Two more solutions of the on/off problem**

Second Solution of the On/Off Problem

Consider all the data at once; the likelihood is a product of Poisson distributions for the on- and off-source counts:

$$\begin{aligned}\mathcal{L}(s,b) &\equiv p(N_{\rm on}, N_{\rm off}|s,b,I) \\ &\propto [(s+b)T_{\rm on}]^{N_{\rm on}} e^{-(s+b)T_{\rm on}} \times (bT_{\rm off})^{N_{\rm off}} e^{-bT_{\rm off}}\end{aligned}$$

Take joint prior to be flat; find the joint posterior and marginalize over b;

$$\begin{aligned}p(s|N_{\rm on}, I_{\rm on}) &= \int db\; p(s,b|I)\,\mathcal{L}(s,b) \\ &\propto \int db\; (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}\end{aligned}$$

$\rightarrow$ same result as before.

Third Solution: Data Augmentation

Suppose we knew the number of on-source counts that are from the background, N_b. Then the on-source likelihood is simple:

$$p(N_{\rm on}|s, N_b, I_{\rm all}) = {\rm Pois}(N_{\rm on} - N_b; sT_{\rm on}) = \frac{(sT_{\rm on})^{N_{\rm on}-N_b}}{(N_{\rm on} - N_b)!} e^{-sT_{\rm on}}$$

Data augmentation: Pretend you have the "missing data," then marginalize to account for its uncertainty:

$$\begin{aligned} p(N_{\rm on}|s, I_{\rm all}) &= \sum_{N_b=0}^{N_{\rm on}} p(N_b|I_{\rm all})\, p(N_{\rm on}|s, N_b, I_{\rm all}) \\ &= \sum_{N_b} \text{Predictive for } N_b \times {\rm Pois}(N_{\rm on} - N_b; sT_{\rm on}) \end{aligned}$$

$$\begin{aligned} p(N_b|I_{\rm all}) &= \int db\; p(b|N_{\rm off}, I_{\rm off})\, p(N_b|b, I_{\rm on}) \\ &= \int db\; {\rm Gamma}(b) \times {\rm Pois}(N_b; bT_{\rm on}) \end{aligned}$$

$\rightarrow$ same result as before.

A profound consistency

We solved the on/off problem in multiple ways, always finding the same final results.

This reflects something fundamental about Bayesian inference.

R. T. Cox proposed two necessary conditions for a quantification of uncertainty:

- It should duplicate deductive logic when there is no uncertainty
- Different decompositions of arguments should produce the same final quantifications (internal consistency)

Great surprise: These conditions are *sufficient*; they lead to the probability axioms. E. T. Jaynes and others refined and simplified Cox's analysis.

Multibin On/Off

The more typical on/off scenario:

Data = spectrum or image with counts in many bins

Model M gives signal rate $s_k(\theta)$ in bin k, parameters θ

To infer θ, we need the likelihood:

$$\mathcal{L}(\theta) = \prod_k p(N_{\mathrm{on}\,k}, N_{\mathrm{off}\,k} | s_k(\theta), M)$$

For each k, we have an on/off problem as before, only we just need the marginal likelihood for s_k (not the posterior). The same C_i coefficients arise.

XSPEC and CIAO/Sherpa provide this as an option.

CHASC approach does the same thing via data augmentation.