Introduction to Bayesian inference: Supplemental topics

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

IAC Winter School, 3–4 Nov 2014

Supplemental topics

1. Binary classification and case diagrams
2. Negative binomial, likelihood principle, prob. & freq.
3. Student's t distribution via marginalization
4. On/off problem: Two more solutions, multibin
5. Parametric bootstrapping vs. posterior sampling

Supplemental topics

1. **Binary classification and case diagrams**
2. Negative binomial, likelihood principle, prob. & freq.
3. Student's t distribution via marginalization
4. On/off problem: Two more solutions, multibin
5. Parametric bootstrapping vs. posterior sampling

Visualizing Bayesian Inference

Simplest case: Binary classification

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

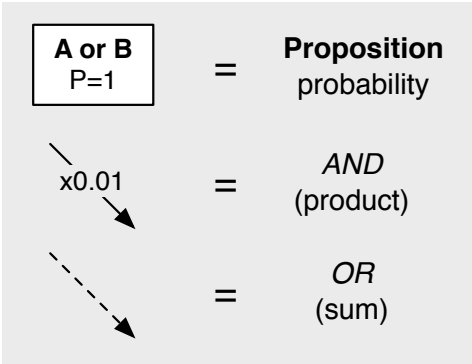
Concrete example: You test positive $(+)$ for a medical condition. Do you have the condition (C) or not $(H$, "healthy"$)$?

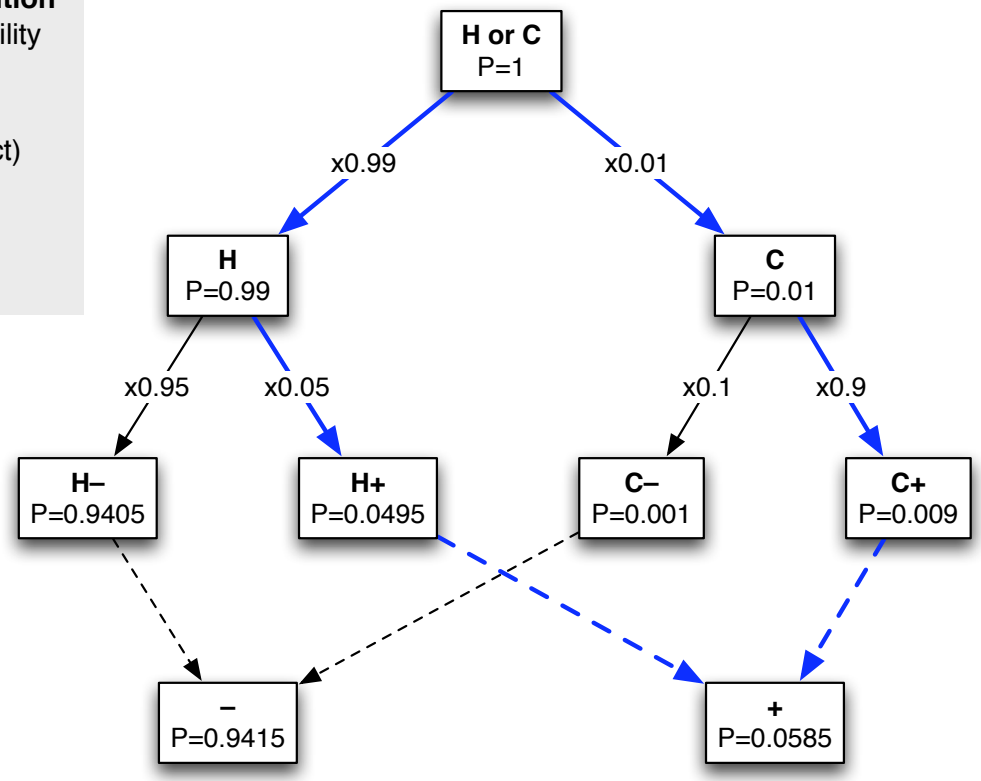
- Prior: Prevalence of the condition in your population is 1%
- Likelihood:
  - Test is 90% accurate if you have the condition: $P(+|C, I) = 0.9$ ("sensitivity")
  - Test is 95% accurate if you are healthy: $P(-|H, I) = 0.95$ ("specificity")

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

[Gigerenzer, *Calculated Risks* (2002)]

Probability "Tree"





$P(H_1 \vee H_2|I)$

$P(H_i|I)$

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

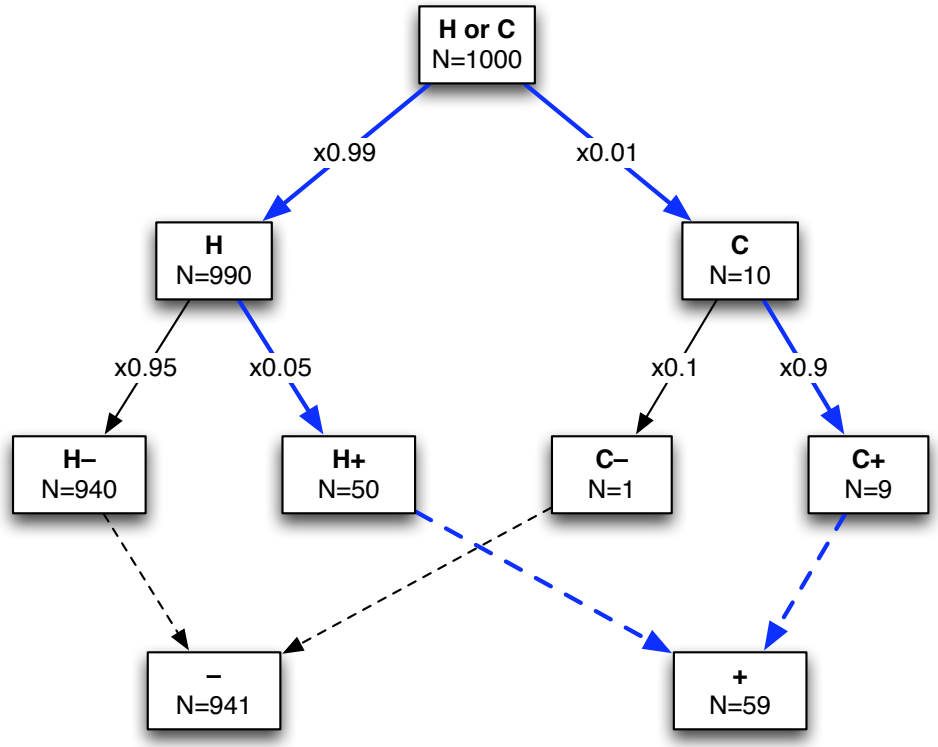
$P(D|I) = \sum_i P(H_i, D|I)$

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

*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. Binary classification and case diagrams
2. **Negative binomial, likelihood principle, prob. & freq.**
3. Student's t distribution via marginalization
4. On/off problem: Two more solutions, multibin
5. Parametric bootstrapping vs. posterior sampling

Another Variation: Negative Binomial

Suppose $D = N$, the number of trials it took to obtain a predefined 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 (well, not rejected) wrt conventional 5% critical level.

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. Binary classification and case diagrams
2. Negative binomial, likelihood principle, prob. & freq.
3. **Student's t distribution via marginalization**
4. On/off problem: Two more solutions, multibin
5. Parametric bootstrapping vs. posterior sampling

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

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

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 $\log \sigma$)

This is also the reference prior, and the minimal sample size prior—the posterior is improper in σ unless $N \geq 2$

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:

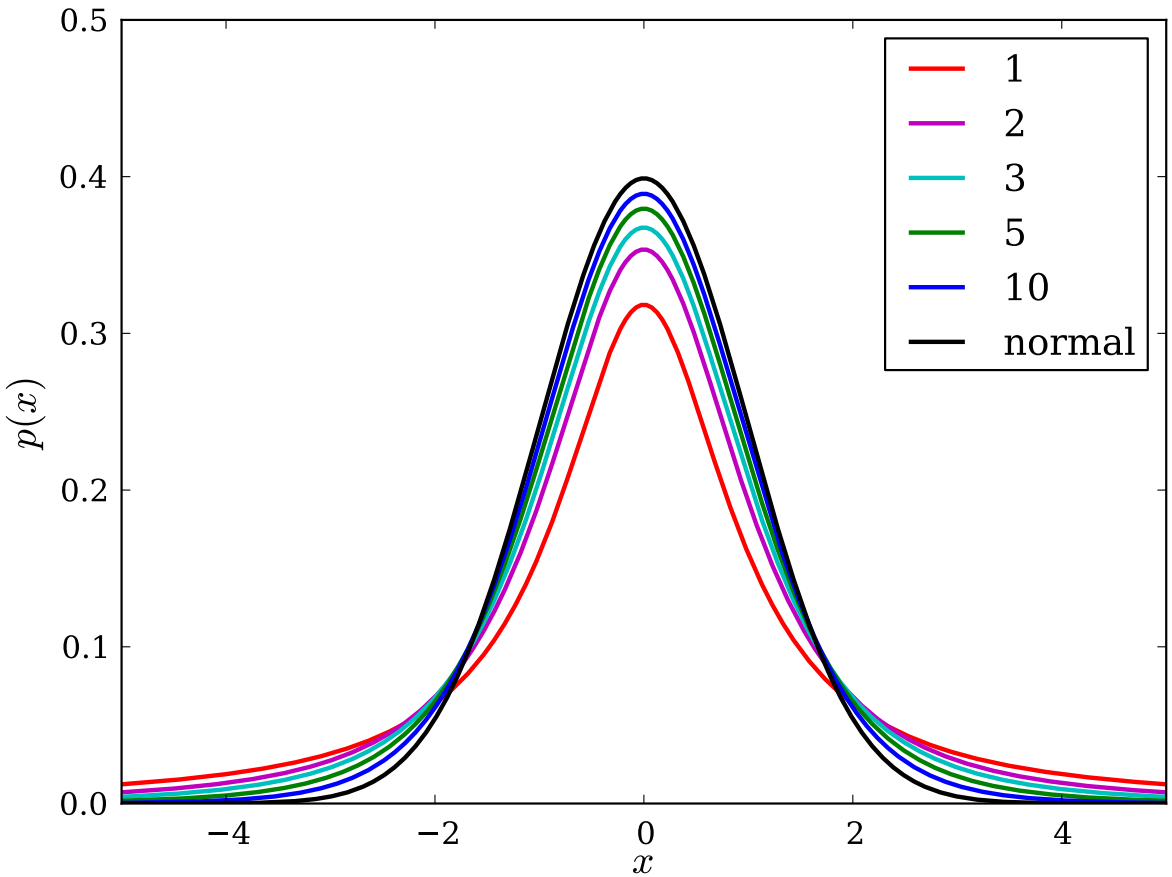
$$p(\mu|D,M) \sim e^{-N(\mu-\overline{d})^2/2r^2}$$

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. Binary classification and case diagrams
2. Negative binomial, likelihood principle, prob. & freq.
3. Student's t distribution via marginalization
4. **On/off problem: Two more solutions, multibin**
5. Parametric bootstrapping vs. posterior sampling

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 (van Dyk$^+$ 2001) does the same thing via Monte Carlo data augmentation

Supplemental topics

1. Binary classification and case diagrams
2. Negative binomial, likelihood principle, prob. & freq.
3. Student's t distribution via marginalization
4. On/off problem: Two more solutions, multibin
5. **Parametric bootstrapping vs. posterior sampling**

Bootstrapping vs. posterior sampling

"Bootstrapping" is a a framework that aims to improve simple but approximate frequentist methods:

- *Parametric bootstrap*: Improve asymptotic behavior of estimates for a trusted model: reduce bias of estimates, provide more accurate coverage of confidence regions
- *Nonparametric bootstrap*: Provide results that are approximately accurate with weak modeling assumptions

Most common approach uses Monte Carlo to simulate an ensemble of data sets related to the observed one, and use them to recalibrate a simple method.

Parametric bootstrap has a step producing an ensemble of estimates that looks like a set of posterior samples. Can they be thought of this way?

Coverage and Confidence Intervals

Setup

A distribution with parameters θ produces data D.
$\theta^* =$ true value of parameters producing many replicate datasets
$D_{\text{obs}} =$ a single, actually observed dataset

Terminology

"Statistic" $\equiv$ Function of data, $f(D)$ (i.e., θ doesn't appear)

"Interval" $\equiv$ Interval-valued statistic $\Delta(D)$, e.g., for 1-D parameter,

$$\Delta(D) = [l(D), u(D)]$$

Note "interval" refers both to the *statistic* (function), and to a *particular interval*, e.g., $\Delta(D_{\text{obs}})$.

Examples:

- Interval about the mean: $\Delta(D) = [\bar{x} - C, \bar{x} + C]$
- Order-statistic-based interval: $\Delta(D) = [x_{(6)}, x_{(11)}]$

"Coverage" $\equiv$ Fraction of time interval contains θ:

$$C(\theta) \;\; = \;\; \int dD \; p(D|\theta) \; [\![\theta \in \Delta(D)]\!]$$

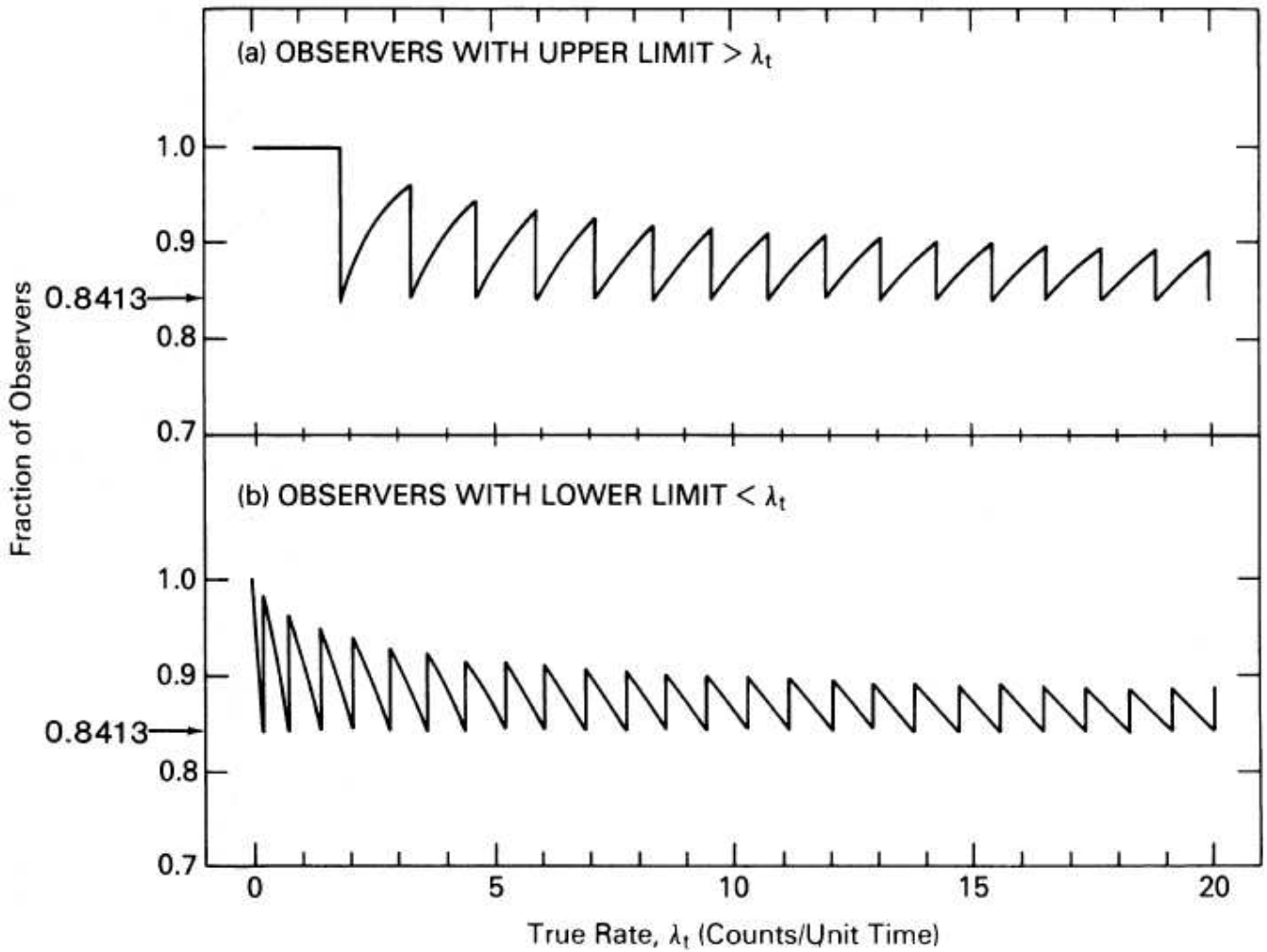
Monte Carlo algorithm using N simulated datasets:

$$C(\theta) \approx \frac{1}{N} \sum_{D \sim p(D|\theta)} [\![\theta \in \Delta(D)]\!]$$

1. Fix θ at some value; start a counter $n = 0$
2. Simulate a dataset from $p(D|\theta)$
3. Calculate $\Delta(D)$; increment counter if $\theta \in \Delta(D)$
4. Goto (2) for N total iterations
5. Report $C(\theta) = \frac{n}{N}$

In general the coverage *depends on* θ.

Coverage of "1σ" upper and lower limits (i.e., 84% confidence level) for a Poisson rate, as a function of the (unknown!) true rate.



Gehrels (1986)

'Plug-In" Approximation

Problem: We don't know θ^* (that's why we're doing statistics!). When we report $\Delta(D_{\text{obs}})$, what coverage should we report?

"Confidence level" CL $\equiv$ maximum coverage over all possible values of θ, a conservative promise of coverage

For complex models, calculating $C(\theta)$ across the whole parameter space is prohibitive.

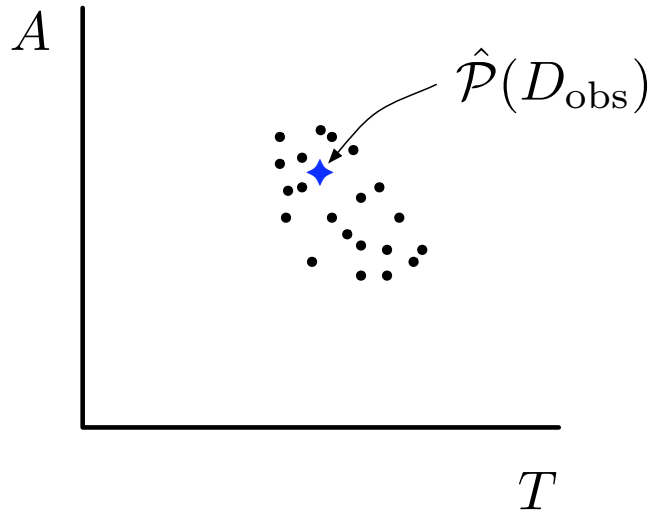
"Plug-in" approach

- Devise some estimator (a statistic!) $\hat{\theta}(D)$ for the parameters; e.g., maximum likelihood
- Calculate $\hat{C} = C(\hat{\theta}(D_{\text{obs}}))$
- Report $\Delta(D_{\text{obs}})$ with $CL \approx \hat{C}$

This gives a *parametric bootstrap* confidence interval; the term is most common when Monte Carlo simulated datasets from $p(D|\hat{\theta}(D_{\text{obs}}))$ are used to estimate $\hat{C}$.

Incorrect Parametric Bootstrapping

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



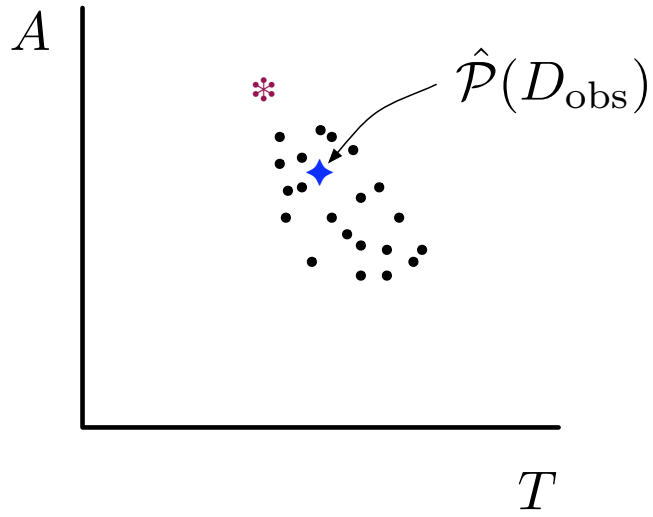
Dots show estimates found by analyzing bootstrapped data sets.

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

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



Dots show estimates found by analyzing bootstrapped data sets.

Histograms/contours of best-fit estimates from $D \sim p(D|\hat{\theta}(D_{\rm obs}))$ 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*.

Likelihood-Based Parametric Bootstrapping

Key idea: Use likelihood *ratios* to define confidence regions.
I.e., use $L = \ln \mathcal{L}$ or χ^2 *differences* to define regions.

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

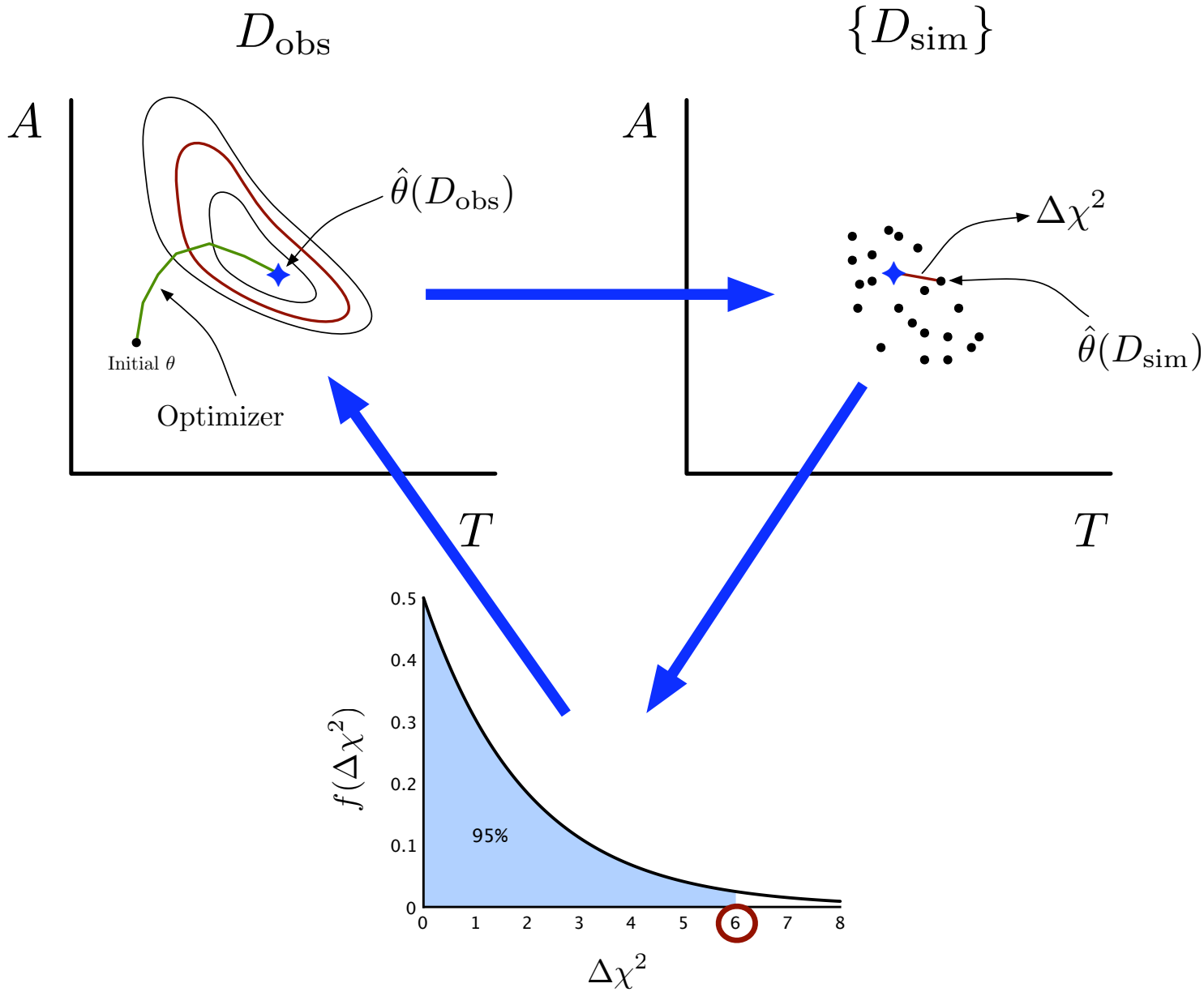
$$\Delta(D) = \{\theta : L(\theta) > L_{\max} - \Delta L\}$$

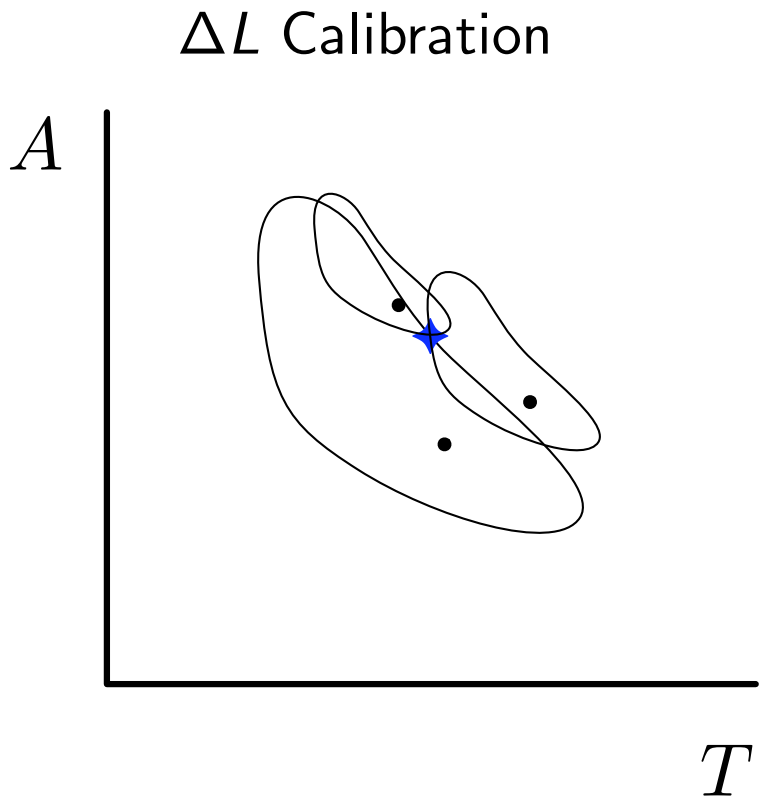
Coverage calculation:

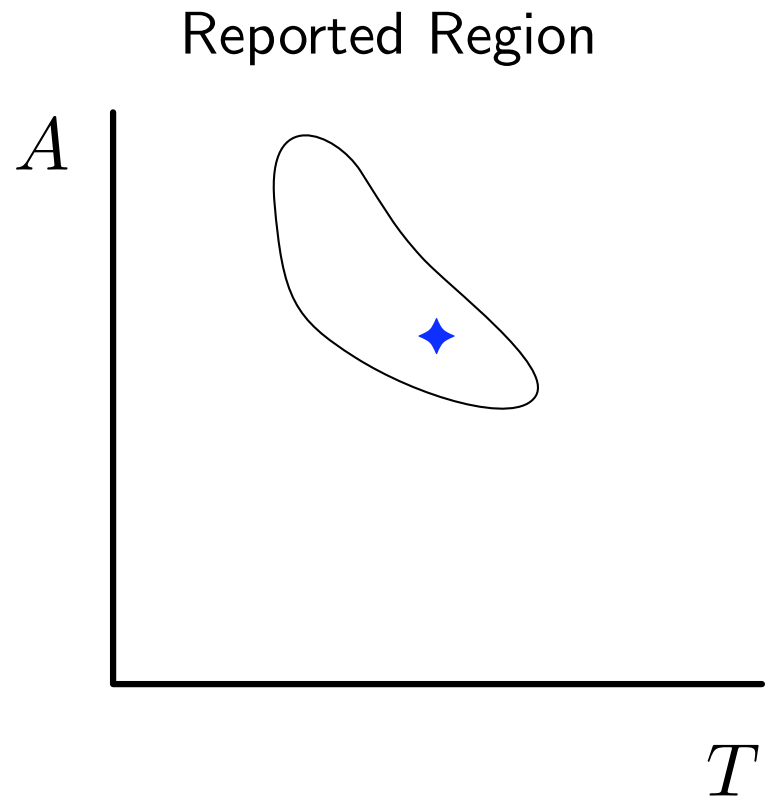
1. Fix $\theta_0 = \hat{\theta}(D_{\text{obs}})$ (plug-in approx'n)
2. Simulate a dataset from $p(D|\theta_0) \to L_D(\theta)$
3. Find maximum likelihood estimate $\hat{\theta}(D)$
4. Calculate $\Delta L = L_D(\hat{\theta}_D) - L_D(\theta_0)$
5. Goto (2) for N total iterations
6. Histogram the ΔL values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

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

Note that CL is a property of the *function* $\Delta(D)$, not of the particular interval, $\Delta(D_{\text{obs}})$.


D_{obs}
$\{D_{\mathrm{sim}}\}$
A
T
$\hat{\theta}(D_{\mathrm{obs}})$
Initial θ
Optimizer
$\Delta\chi^2$
$\hat{\theta}(D_{\mathrm{sim}})$
$f(\Delta\chi^2)$
95%
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8
$\Delta\chi^2$





The CL is approximate due to:

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

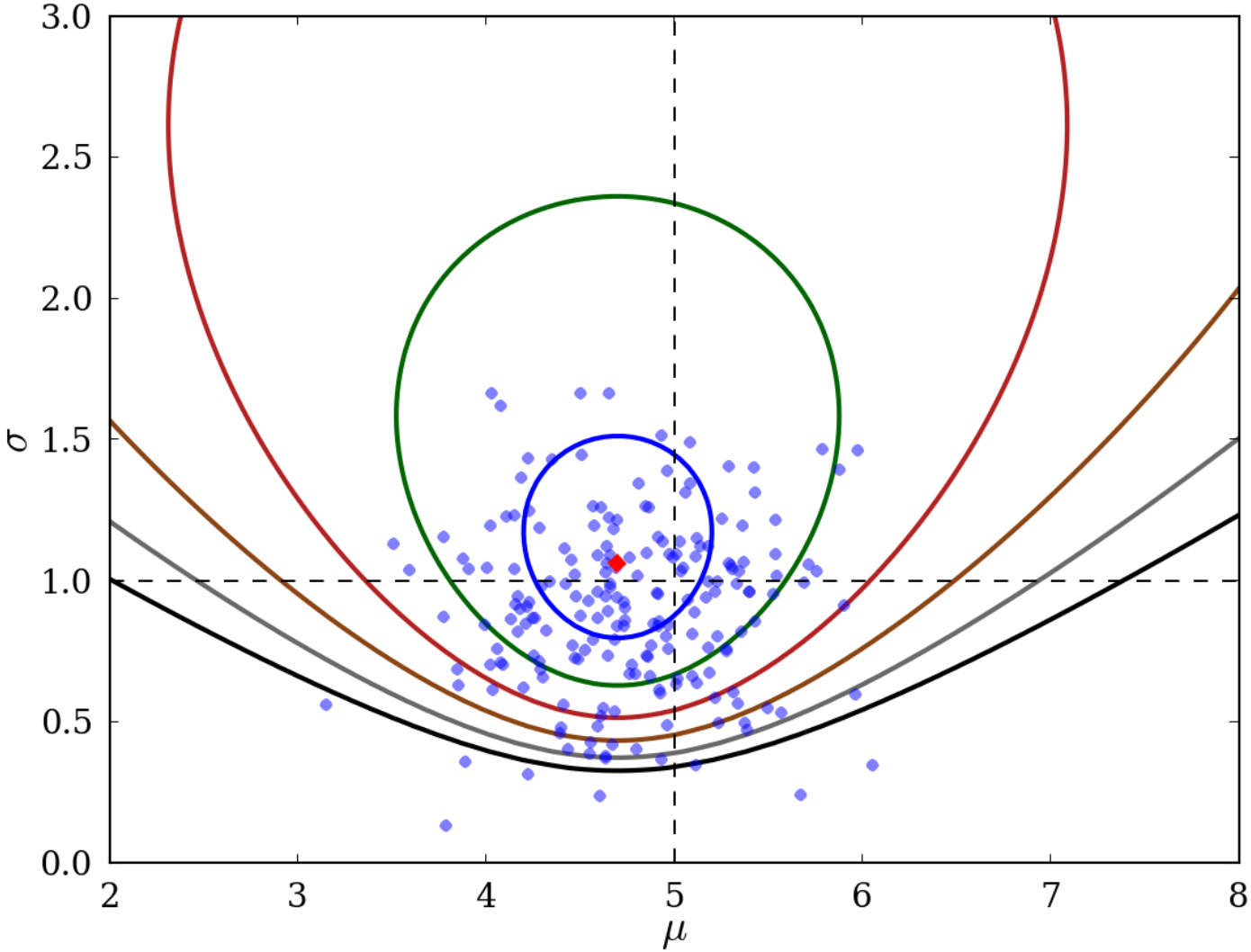
Skewness/asymmetry in the distribution of bootstrap estimates is the rule rather than the exception!

Simple example: Estimate the mean *and standard deviation* of a normal distribution

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

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

Results for $\mu = 5$, $\sigma = 1$, $N = 5$; 200 samples:



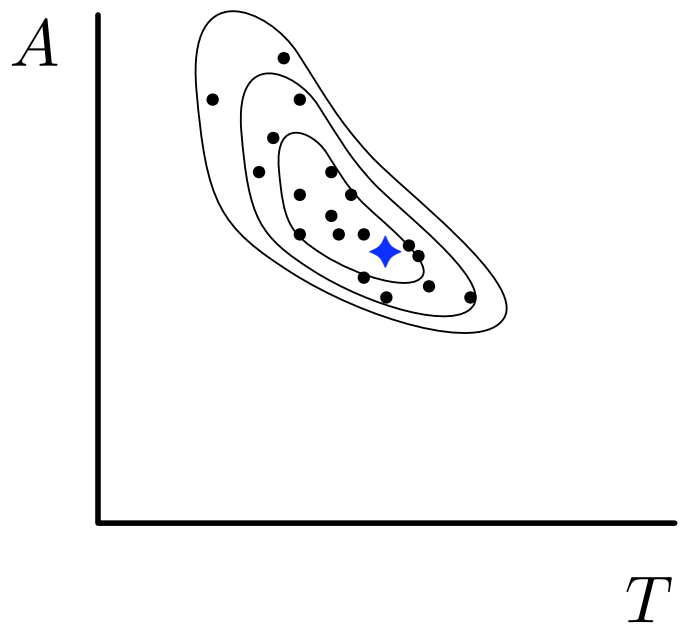
Points are skewed *down*, so the truth is most likely *up*—as indicated by the likelihood contours

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.



Interpretations of Regions

Confidence region

Frequentist probabilities describe *variability* in the performance of procedures/rules over an ensemble.

A confidence region $\Delta(D)$ with specified CL contains the true parameter value 100CL% of the time. This quantifies the confidence you might have that the value is in the particular interval $\Delta(D_{\text{obs}})$.

Credible region

Bayesian probabilities are quantifications of the *strength of arguments*—$p(A|B)$ measures how justified one is in reasoning from B to A, i.e., how strongly B supports A vs. its alternatives.

The probability within a credible region quantifies how strongly the particular dataset we've observed justifies concluding the true parameter value is in the region.

Performance of credible regions

How often may we expect an HPD region to include the true value if we analyze many datasets? I.e., what's the performance of an interval rule $\Delta(D)$ defined by calculating the HPD region each time?

Suppose we generate datasets by picking a parameter value from $\pi(\theta)$ and simulating data from $p(D|\theta)$.

The fraction of time θ will be in the HPD region is:

$$Q = \int d\theta \, \pi(\theta) \int dD \, p(D|\theta) \, [\![\theta \in \Delta(D)]\!]$$

Note $\pi(\theta)p(D|\theta) = p(\theta, D) = p(D)p(\theta|D)$, so

$$Q = \int dD \int d\theta \, p(\theta|D) \, p(D) \, [\![\theta \in \Delta(D)]\!]$$

$$
\begin{aligned}
Q &= \int dD \int d\theta \; p(\theta|D)\, p(D) \; [\![\theta \in \Delta(D)]\!] \\
&= \int dD \; p(D) \int d\theta \; p(\theta|D) \; [\![\theta \in \Delta(D)]\!] \\
&= \int dD \; p(D) \int_{\Delta(D)} d\theta \; p(\theta|D) \\
&= \int dD \; p(D) P \\
&= P
\end{aligned}
$$

The HPD region includes the true parameters $100P\%$ of the time.

This is exactly true for any problem, even for small datasets.

Keep in mind it involves drawing θ from the prior; credible regions are "calibrated with respect to the prior."

Average Coverage

Recall the original integral:

$$\begin{aligned} Q &= \int d\theta\, \pi(\theta) \int dD\; p(D|\theta)\; [\![\theta \in \Delta(D)]\!] \\ &= \int d\theta\, \pi(\theta) C(\theta) \end{aligned}$$

where $C(\theta)$ is the coverage of the HPD region when the data are generated using θ.

This indicates Bayesian regions have accurate *average coverage*.

The prior can be interpreted as quantifying how much we care about coverage in different parts of the parameter space.

Frequentist Performance of Bayesian Procedures

Many results known for parametric Bayes performance:

- Estimates are consistent if the prior doesn't exclude the true value.
- Credible regions found with flat priors are typically confidence regions to $O(n^{-1/2})$ (Bernstein-von Mises Theorem); "reference" priors can improve their performance to $O(n^{-1})$.
- Marginal distributions have better frequentist performance than conventional methods like profile likelihood. (Bartlett correction, ancillaries are competitive but hard.)
- Bayesian model comparison is asymptotically consistent (not true of significance/NP tests, AIC).
- Misspecification: Bayes converges to the model with sampling dist'n closest to truth via Kullback-Leibler

- Results are just appearing for nonparametric & semiparametric models; *you must be more careful with priors here*
- Wald's complete class theorem: *Optimal* frequentist methods are *Bayes rules* (equivalent to Bayes for some prior)
- . . .

Parametric Bayesian methods are typically good frequentist methods.

Some references:

- "The Interplay of Bayesian and Frequentist Analysis" (Bayarri & Berger 2004) *Statistical Science*, **19**, 58–80
- "Calibrated Bayes: A Bayes/Frequentist Roadmap" (Little 2006; 2005 ASA President's Invited Address) *The American Statistician*, **60**, 213–223

Rescuing the bootstrap parameter estimates

Although the best-fit parameters from bootstrapped data don't correspond to posterior samples, they are in the neighborhood of the posterior $\rightarrow$ use them to create an importance sampling distribution:

- Weighted Likelihood Bootstrap: Nonparametric bootstrap + KDE for modest-dimensional models (Newton & Raftery 1994)
- Efron (2011, 2012): Posterior sampling via parametric bootstrap and importance sampling adjustments