Bayesian Statistical Inference

A Tuesday in 2024 Winter

9:50-12:15

三教1200

清華大學天文系

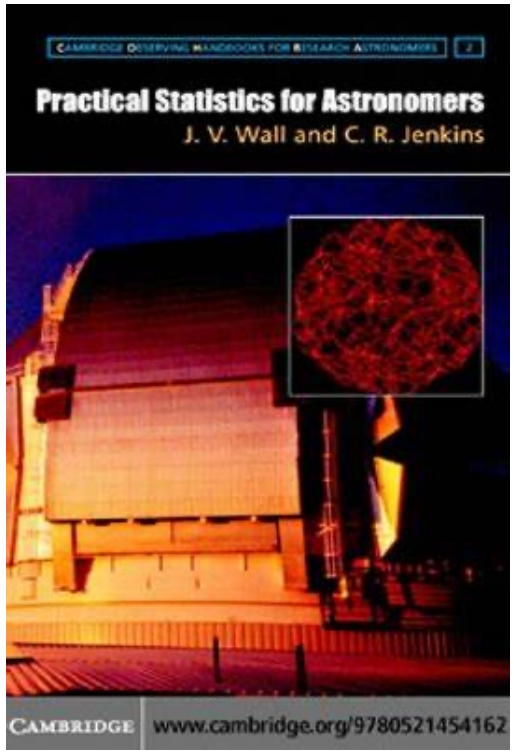
Department of Astronomy, Tsinghua University

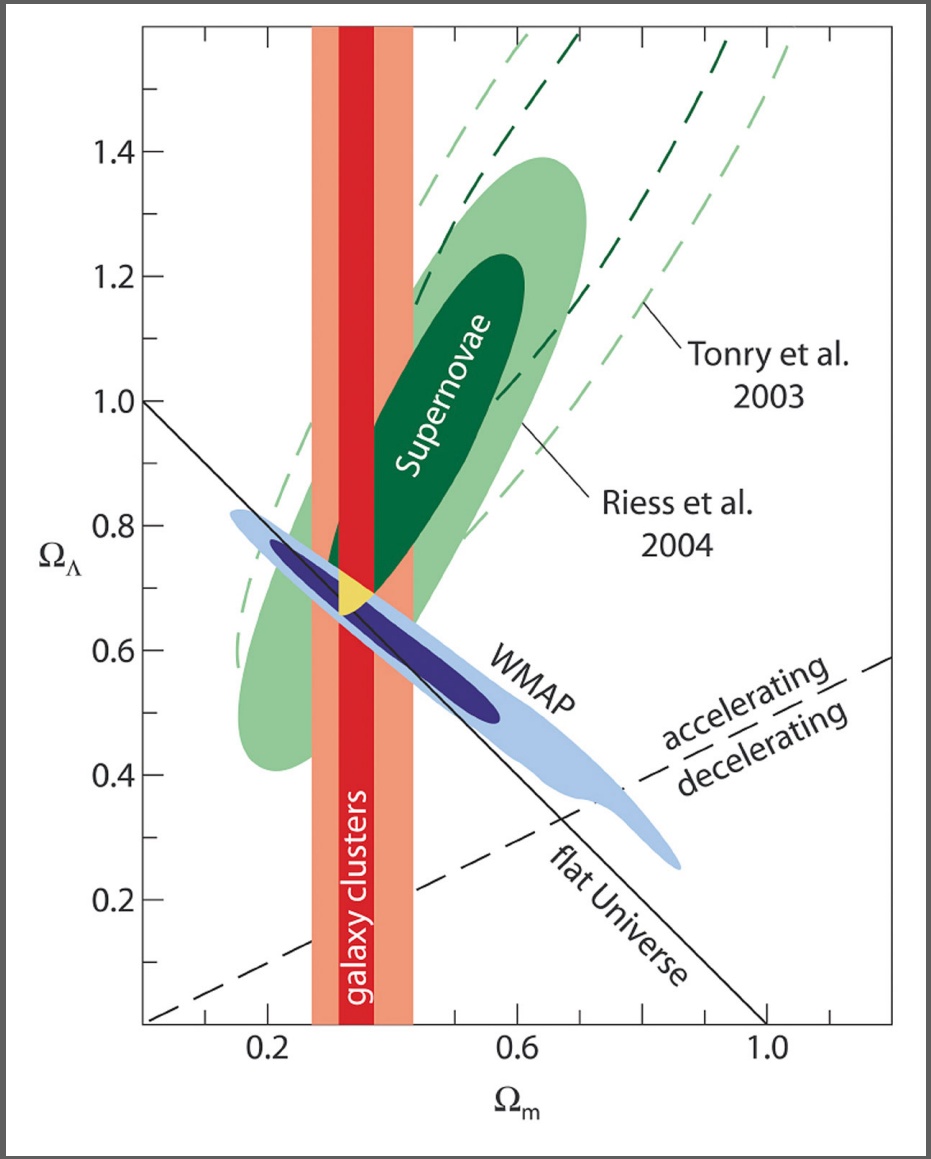
Classical Statistical Inference

"The application of efficient statistical procedure has power;

but the application of common sense has more!"

Wall and Jenkins





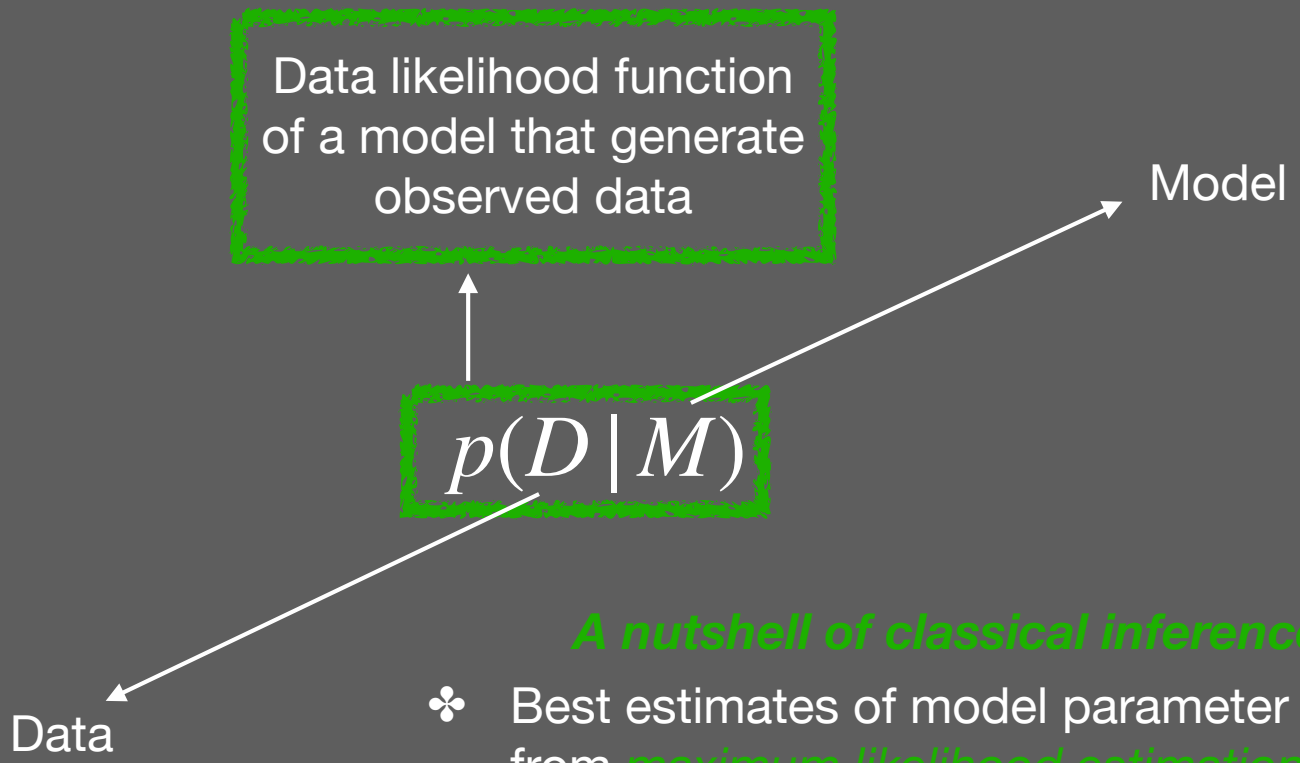
Parameter estimation is a very basic task in modern sciences. Frequentist statistics has been gradually replaced by Bayesian statistics.

1. Basic formula
2. Essential ingredients
3. Applications

1. Bayesian Framework

What is Bayesian statistical inference?

Classical statistical inference
"a frequentist approach"



A nutshell of classical inference:

- Best estimates of model parameter come from *maximum likelihood estimation* (MLE)
- Uncertainty estimates of model parameter come from Fisher information matrix or through *bootstrap/jackknife.*

In a nutshell

For any event A, B, the joint probability:

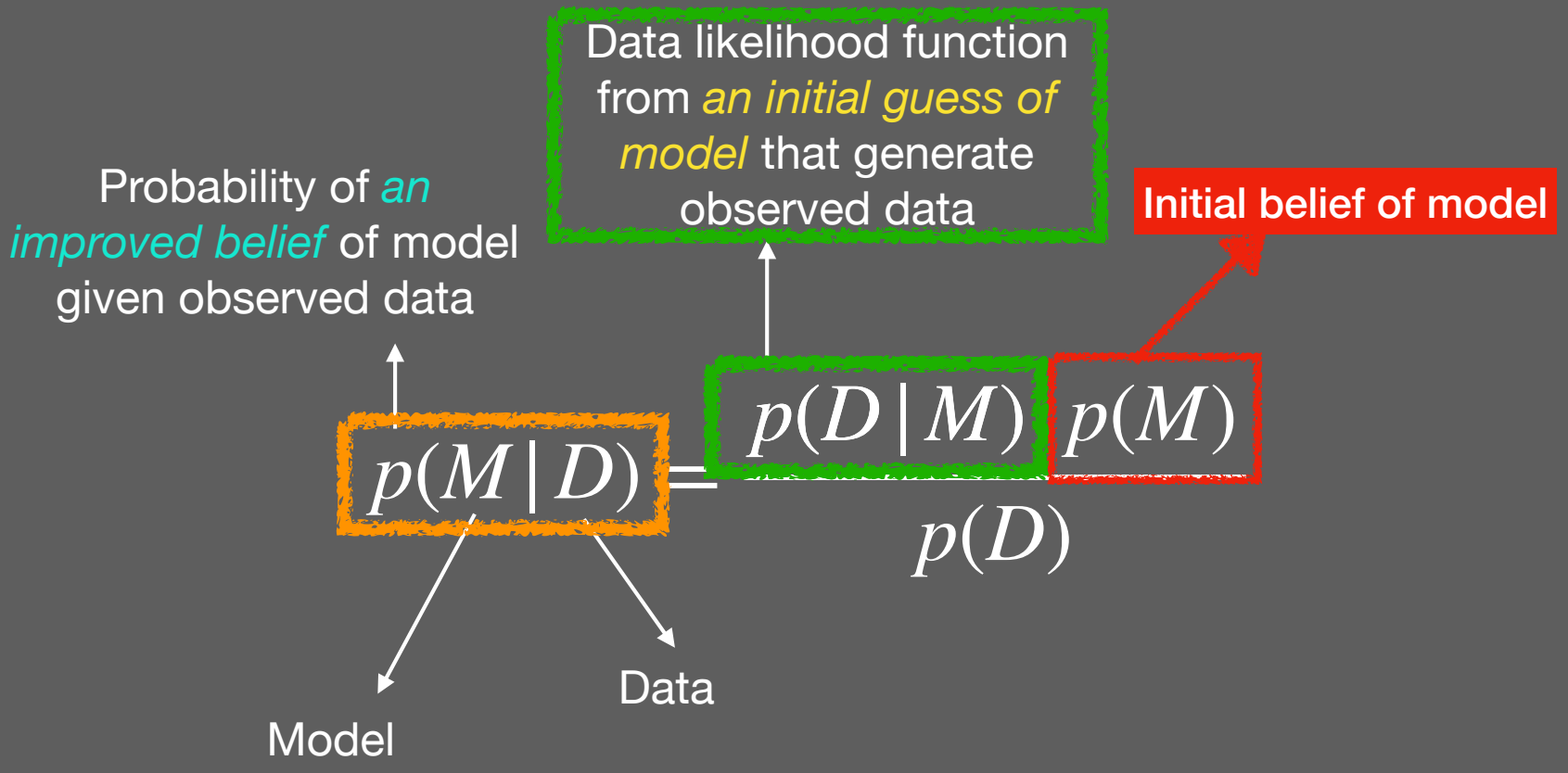
$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

Symmetrize the probability calculation between data and model/parameter!

The conditional probability:

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)}$$

Bayesian Formula



In a nutshell

Bayes' theorem full formula

The Prior *Parameter* *Model*

$$p(M, \vec{\theta} \mid I) = p(\vec{\theta} \mid M, I)\ p(M \mid I)$$

Prior - "a priori joint probability" given *prior information* I, and in the absence of any data D!

Prior: old knowledge/ constraints on model

Posterior pdf of model $M[\theta]$ given observed data and prior information

Likelihood function of model $M[\theta]$ that generate observed data D, under the *prior information* I

Posterior: + new knowledge/ constraints on model

$$p(M, \vec{\theta} \mid D, I) = \frac{p(D \mid M, \vec{\theta}, I)\ p(M, \vec{\theta} \mid I)}{p(D \mid I)}$$

Model M described by parameter θ

Given data and *prior information*

Move from old knowledge/constraints on model to a new knowledge/constraints

In a nutshell

What does Bayesian statistical inference do?

1. Model $M[\overrightarrow{\theta}]$ is given and we are doing parameter $\overrightarrow{\theta}$ estimation within a Bayesian framework

2. Considering model selection, i.e., comparing between models, within a Bayesian framework

In a nutshell

1. Model $M[\vec{\theta}]$ is given and we are doing parameter $\vec{\theta}$ estimation within a Bayesian framework

$$p(\vec{\theta} \mid D, M, I) = \frac{p(D \mid \vec{\theta}, M, I)\, p(\vec{\theta} \mid M, I)}{p(D \mid M, I)}$$

For a given model M, prior normalization: $\int p(\vec{\theta} \mid M, I)\, \mathrm{d}\vec{\theta} = 1$

Posterior pdf normalization meaning: $\int p(\vec{\theta} \mid D, M, I)\, \mathrm{d}\vec{\theta} = 1$

It requires that $p(D \mid M, I) = \int p(D \mid \vec{\theta}, M, I) p(\vec{\theta} \mid M, I)\, \mathrm{d}\vec{\theta}$

$E(M) \equiv p(D \mid M, I)$ is also called the *evidence* (model likelihood).

Model evidence is used as normalization-keeping at parameter estimate level, given model.

In a nutshell

Once we obtain the posterior $p(\overrightarrow{\theta}\,|D,M,I) = p(\theta_1,\theta_2,\theta_3\ldots,\theta_k\,|D,M,I)$ for a given model $M(\overrightarrow{\theta})$ under prior information I, we can further calculate:

Marginalization of Parameters:

nuisance parameters

★ If we are interested in posterior pdf of θ_1:

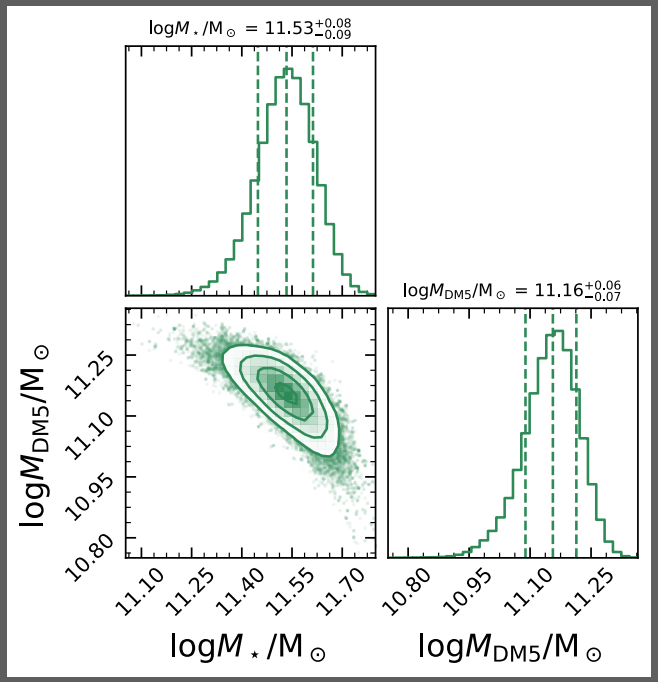
$$p(\theta_1\,|D) = \int p(\theta_1,\theta_2,\theta_3\ldots,\theta_k\,|D)\,\mathrm{d}\theta_2\,\mathrm{d}\theta_3\ldots\mathrm{d}\theta_k$$

★ If we are interested in covariance between θ_1 and θ_2:

$$p(\theta_1,\theta_2\,|D) = \int p(\theta_1,\theta_2,\theta_3\ldots,\theta_k\,|D)\,\mathrm{d}\theta_3\ldots\mathrm{d}\theta_k$$

marginal posterior pdf *multidimensional posterior pdf*

Some can be real model parameters and some can be hyperparameters



Posterior Interval - credible region (a,b) of parameter θ:

Once we obtain the posterior pdf of θ, we can estimate a Bayesian credible region by finding a and b, such that $\int_{-\infty}^{a} p(\theta)\,\mathrm{d}\theta = \int_{b}^{\infty} p(\theta)\,\mathrm{d}\theta = \alpha/2$. Then (a,b) is called $1-\alpha$ *posterior interval*. In practice, the posterior pdf $p(\theta)$ is often not an analytical function, finding $1-\alpha$ posterior interval of θ can be done via Monte Carlo sampling.

In a nutshell

To find best model parameters:
Maximum likelihood estimate (MLE)

To quantify parameter uncertainties:
— Confidence regions:

1. Through Fisher Information matrix
2. Bootstrap or Jackknife (resampling)

Classical/frequentist perspective

Bayesian

To find best model parameters:

Maximize either the joint posterior $p(\vec{\theta} \mid D)$ or the marginalized posterior $p(\theta \mid D)$ to yield the *maximum a posteriori (MAP)* estimate

Or

Obtain the *posterior mean (or median)*:

$$\bar{\theta}_1 = \int \theta_1 \, p(\theta_1 \mid D) \, \mathrm{d}\theta_1,$$

where $p(\theta_1 \mid D)$ is obtained using marginalization - integration of $p(\vec{\theta} \mid D)$ over all other model parameters and don't forget to renormalize: $\int p(\theta_1 \mid D) \, \mathrm{d}\theta_1 = 1$

To quantify parameter uncertainties
— Credible regions:

Analytically calculating the (normalized) posterior interval or sampling the posterior distribution numerically (e.g., MCMC).

In a nutshell

2. Considering model selection, i.e., comparing between models, within a Bayesian framework

$$p(M \mid D, I) = \frac{p(D \mid M, I)\, p(M \mid I)}{p(D \mid I)}$$

$$E(M) \equiv p(D \mid M, I)$$

Model evidence is the likelihood at model estimate level!

Normalization of posterior, i.e., $\int p(M \mid D, I)\, \mathrm{d}M = 1$, requests:

$$p(D \mid I) = \int p(D \mid M, I) p(M \mid I)\, \mathrm{d}M$$

Probability of data,

Prior predictive probability for D,

Normalization-keeping

Information criteria or the odds ratio based on Bayesian likelihood or posterior can then be used be decide which model is better.

In a nutshell

2. Essential ingredients

2.1 likelihood (same as classical inference)

e.g., Gaussian likelihood (large number statistics),
Poisson likelihood (rare events, small number statistics),
Binomial likelihood (binary choice, finite # of events)

2.2 Priors (unique feature of Bayesian inference)!

I. Uninformative/weak priors
II. Conjugate priors
III. Hierarchical priors

Bayesian Priors $p(\vec{\theta} \mid M, I)$

Given model

Priors have nothing to do with the current dataset that the model will be implemented to, i.e., dataset *D* does not get involved in the prior calculation.

But priors can come from some posterior distributions that were achieved before using previous dataset in the previous analysis.

In a way, the inclusion/implementation of priors makes the Bayesian inference a refining process from previous knowledge/ belief to new improved knowledge/belief.

e.g., when estimating cosmological parameters using supernova observations, the prior may comes from CMB and LSS measurements etc.

I. Uninformative/weak priors
II. Conjugate priors
III. Hierarchical priors

Uninformative/weak priors

Bayesian Priors - I: Uninformative/weak priors

e.g. "the model parameter describing variance cannot be negative."

Despite of "uninformative", it can still affect the estimate and the results are not generally equivalent to the classical inference results!

- *How to write a uninformative prior $p(\theta)$?*

✦ *Principle of indifference*

A fair six-sided dice, each side has a prior probability of 1/6:
e.g., $p(\theta) = 1/n$

✦ *Principle of consistency - I*

A flat prior for a *location parameter* μ should remain flat under coordinate translation, i.e., the translation invariant condition:
$p(\mu \,|\, I)\,\mathrm{d}\mu = p(\mu + a \,|\, I)\,\mathrm{d}\mu$ requests
$p(\mu) = C$ as a solution for the prior.

✦ *Principle of consistency - II*

A flat prior on a *scale parameter* σ (e.g., standard deviation) $p(\sigma)$ should not depend on the choice of units, i.e., if rescaling measurement units by a, then the constraint of $p(\sigma \,|\, I)\,\mathrm{d}\sigma = p(a\sigma \,|\, I)\,\mathrm{d}(a\sigma)$ shall be satisfied. This requires a solution that $p(\sigma) \propto \sigma^{-1}$ or equivalently a flat prior $p[\ln(\sigma)] = 1/C$
- this is called a scale-invariant prior.

Bayesian Priors - I: Uninformative/weak priors

✦ *Principle of Maximum Entropy:*

Shannon's entropy:

$S = -\Sigma_i^N p_i \log p_i$,

where N is the number of different outcomes.

$\Sigma_i^N p_i = 1$ and $\lim_{p\to 0}[p \ln p] = 0$.

Entropy S has a unit of *nat* for $\log_e$ and *bit* for $\log_2$.

Define the relative entropy:

$S_{\text{rel}} = -\Sigma_i^N p_i \log \left(\frac{p_i}{m_i}\right)$ for discrete case

$S_{\text{rel}} = -\int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{m(x)} \mathrm{d}x$ for continuous case

$m(x)$ can be a distribution without any prior knowledge.

Kullback-Leibler (KL) divergence from $p(x)$ to $m(x)$

Head **Tail**

"Least information carried" by 3 coins is the combination of all *N* possible outcomes, each with an equal probability:

$p_i = 1/N$, where *N* is *the* # of outcomes, $N = 2^3$. The entropy

$S = -\Sigma_i^N \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N$,

so Shannon entropy S = 3 bits.

The KL divergence measures the difference between two pdfs. It is not a true distance metric because the divergence is not the same when *p(x)* and *m(x)* switch. This is often used in measuring information gain when moving from a prior to posterior.

Maximizing entropy method is to maximizing S_{rel}, or minimizing KL divergence from p(x) to some desired (least informative) distribution m(x). This allows us to distribute probabilities among all states with least informative/weak prior!

Bayesian Priors - I: Uninformative/weak priors

Assigning six *prior probabilities* $p_i\,(i = 1, 2, 3, 4, 5, 6)$ to a six-faced dice:

The expected probabilities are $m_i = 1/6$, when not additional information is known.

Now we also know that the dice has a mean value μ for a large number of rolls (e.g., in the case of a fair dice, $\mu = 3.5$).

Now we have two weak constrains:

$$\Sigma_{i=1}^{6} p_i = 1 \quad \Sigma_{i=1}^{6} i\, p_i = \mu \qquad S = -\Sigma_{i=1}^{6} p_i \ln\left(\frac{p_i}{m_i}\right)$$

Maximizing Q with respect to p_i (in combination with the *Lagrangian multiplier* method):

$$Q = S + \lambda_0\left(1 - \Sigma_{i=1}^{6} p_i\right) + \lambda_1\left(\mu - \Sigma_{i=1}^{6} i\, p_i\right)$$

=> *Constraint condition on* p_i*:* $-\left[\ln\left(\frac{p_i}{m_i}\right) + 1\right] - \lambda_0 - i\lambda_1 = 0$

Maximum Entropy Solution:

$$p_i = m_i \exp(-1 - \lambda_0)\, \exp(-i\lambda_1), \text{where } m_i = 1/6.$$

Given p_i, then λ_0 and λ_1 can be determined numerically from the two constraints:

$$\Sigma_{i=1}^{6} p_i = 1 \qquad \Sigma_{i=1}^{6} i\, p_i = \mu$$

Discrete case

Poisson distribution

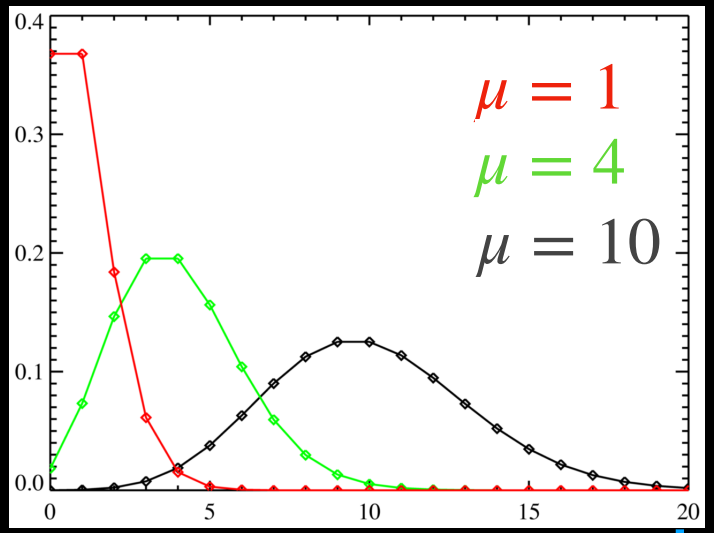
A special case of binomial distribution is when $N \to \infty$ and the success rate b is fixed. The average number of successful event is $\mu = N/b$. The expected # of success (over some time interval) is then given by:

$$p(k \mid \mu) = \frac{\mu^k \exp(-\mu)}{k!}$$

k: Actual number of success

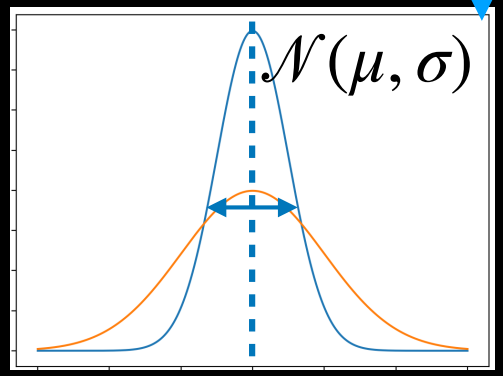
e.g., the number of radioactive decay events,
the number of photons that hit a detector,
the number of patients in a hospital …

$$\bar{k} = \mu \qquad \sigma_k = \sqrt{\mu} \qquad k_{\text{mode}} = \mu - 1$$



$\Sigma = 1/\sqrt{\mu} \to 0$, as $\mu \Uparrow$

$K = 1/\mu \to 0$, as $\mu \Uparrow$



Binomial $p(k \mid b, N)$ $\xrightarrow{N \to \infty}$ Poisson $p(k \mid \mu)$ $\xrightarrow[b \Uparrow]{\mu \Uparrow}$ Gaussian $\mathcal{N}(\mu, \sqrt{\mu})$

Law of small numbers

Law of rare event: b is small

Law of big numbers

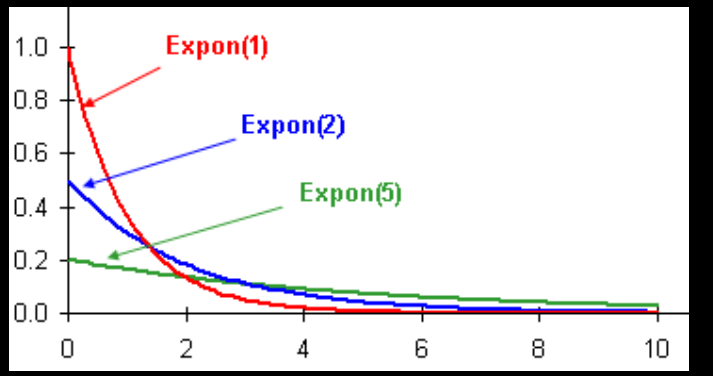
Continuous case

Exponential (Laplace) distribution

$$p(x\,|\,\mu,\Delta) = \frac{1}{2\Delta}\exp\left(\frac{-|x-\mu|}{\Delta}\right)$$

$$\Rightarrow p_{\exp}(t\,|\,\tau) = \tau^{-1}\exp(-t/\tau)$$

$p_{\exp}(x)$



★ One-side exponential, describing the probability of a time span t between two successive and independent events that occur, where τ is the expected average time span between two events, $\bar{t} = \tau$.

Recall: given a time span of t (starting from right after the previous event), the average number of events is $\mu = t/\tau$, then the probability of the number of events happening during this time is given by a ***Poisson distribution***:

$$p_{\text{pois}}(k\,|\,\mu = t/\tau) = \frac{\mu^k\exp(-\mu)}{k!} = \frac{(t/\tau)^k\exp(-t/\tau)}{k!}.$$

★ The probability that new events would occur during this time span t is given by:

$$P^{\geqslant 1}_{\text{Pois}} = 1 - p_{\text{Pois}}(0\,|\,\tau,t) = 1-\exp(-t/\tau), \quad -\textit{ the CDF of exponential;}$$

$$\text{then } \frac{\mathrm{d}P^{\geqslant 1}_{\text{Pois}}}{\mathrm{d}t} = \tau^{-1}\exp(-t/\tau) = p_{\exp}(t\,|\,\tau) - \textit{ the PDF of exponential.}$$

Bayesian Priors - I: Uninformative/weak priors

In the case of discrete outcomes (θ) with infinite possibilities, if the expected number of outcomes μ_0 is *known*, then *maximum entropy solution* for the prior of θ is a Poisson distribution:

μ_0 is the prior information, $p(\theta|\mu_0)$ is the prior!

$$\boxed{p(\theta|\mu_0)} = \frac{\mu^\theta \exp(-\mu)}{\theta!}$$

In the corresponding continuous case for parameter θ, if the mean $\mu_\theta = \int \theta p(\theta) \mathrm{d}\theta$ is *known*, the *maximum entropy solution* for the prior is an exponential distribution:

μ_0 is the prior information, $p(\theta|\mu_0)$ is the prior!

$$\boxed{p(\theta|\mu_0)} = \frac{1}{\mu_0} \exp\left(\frac{-\theta}{\mu_0}\right).$$

And if both the mean $\mu_\theta = \int \theta p(\theta) \mathrm{d}\theta$ and the variance $V_\theta = \int (\theta - \mu)^2 p(\theta) \mathrm{d}\theta$ are *known*, the *maximum entropy solution* for the prior is a *Gaussian* distribution:

μ_0 and V_0 are the *known* prior information, $p(\theta|\mu_0, V_0)$ is the prior!

$$\boxed{p(\theta|\mu_0, V_0)} = \frac{1}{\sqrt{2\pi V_0}} \exp\left(\frac{-(\theta - \mu_0)^2}{2V_0}\right).$$

Conjugate priors

Bayesian Priors - II: Conjugate Priors

When the *posterior* and the *prior pdf* take the same functional form, then the prior is called a conjugate prior. What is the *advantage* of having conjugate priors?

$$p(\vec{\theta} \mid D, I) \propto p(D \mid \vec{\theta}, I)\, p(\vec{\theta} \mid I)$$

New guess on parameter from new data

Old guess on parameter from previous data

The advantage of having the same form of posterior and prior is that the posterior from a previous experiment can be fed to the prior in a new experiment!

This of course depends on the form of likelihood. For specific form of the likelihood function, there are correspondingly matched forms for the conjugate priors, e.g.,

For a Gaussian likelihood, the conjugate prior is a Gaussian distribution.
For a Poisson likelihood, the conjugate prior is a Gamma distribution.
For a binomial likelihood, the conjugate prior is a Beta distribution.

Recall that Poisson and Binomial distributions are also commonly used likelihood functions

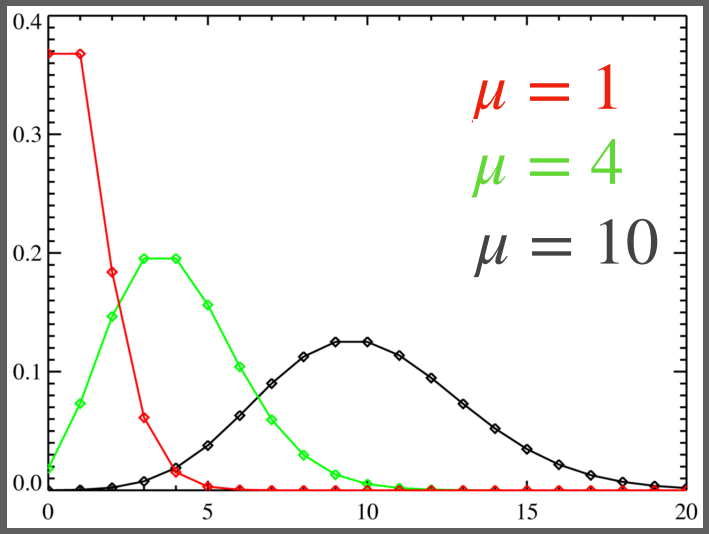
Poisson distribution

A special case of binomial distribution is when $N \to \infty$ and the success rate b is fixed. The average number of successful event is $\mu = N/b$. The expected # of success (over some time interval) is then given by:

$$p(k|\mu) = \frac{\mu^k \exp(-\mu)}{k!}$$

data: k — Actual number of success

parameter: μ



$$\bar{k} = \mu \qquad \sigma_k = \sqrt{\mu} \qquad k_{\text{mode}} = \mu - 1$$

What are the conjugate priors for μ for a Poisson likelihood?

Binomial distributions

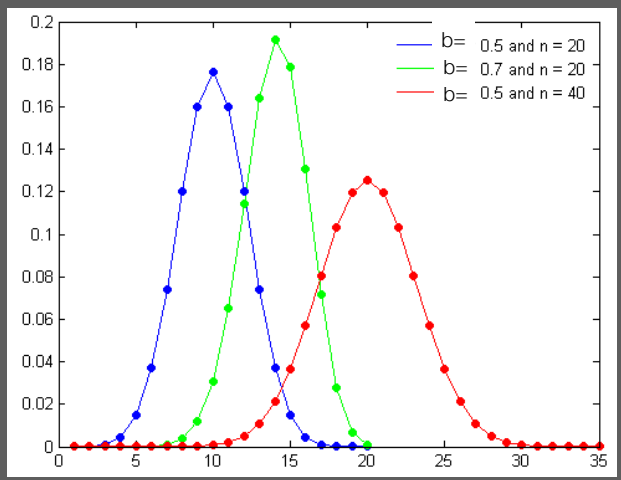
Distributions of the probability of a variable that can only take two discrete values (0, 1), e.g., flip a coin!

$$p(k|b,N) = C_N^k b^k (1-b)^{N-k} = \frac{N!}{(N-k)!k!} b^k (1-b)^{N-k}$$

data: k — # of times success occurred

parameter: b — Probability of success

N — # of total trials



$$\bar{k} = Nb \qquad \sigma_k^2 = Nb(1-b)$$

What are the conjugate priors of b for a binomial likelihood?

Review

$$p(x \mid k, \theta) = \frac{\theta^{-k}}{\Gamma(k)} x^{k-1} \exp(-\frac{x}{\theta})$$

Gamma distribution

parameter: θ; data (old): x

data (new): m; parameter: μ

$$p_{\text{Gamma}}(\mu \mid k = m+1, \theta = 1) \rightarrow p_{\text{Poisson}}(m \mid \mu) = \frac{\mu^m \exp(-\mu)}{m!}$$

$$p(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

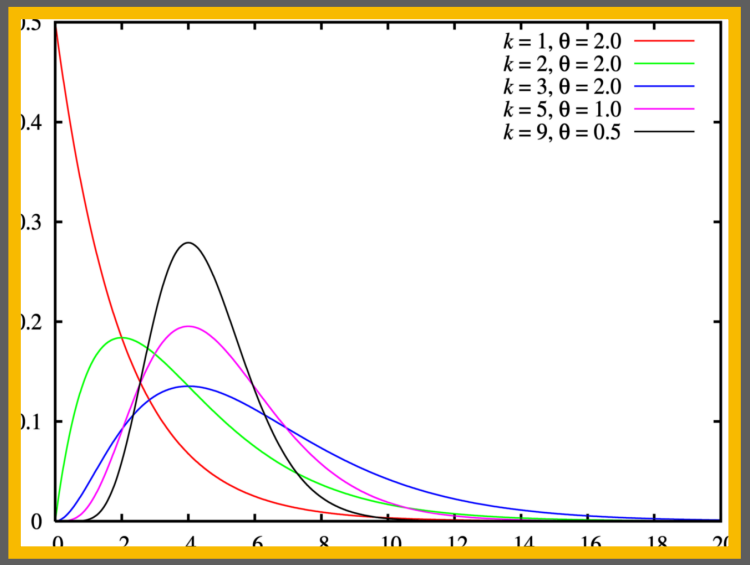
Beta distribution

parameter: α, β; data (old): x

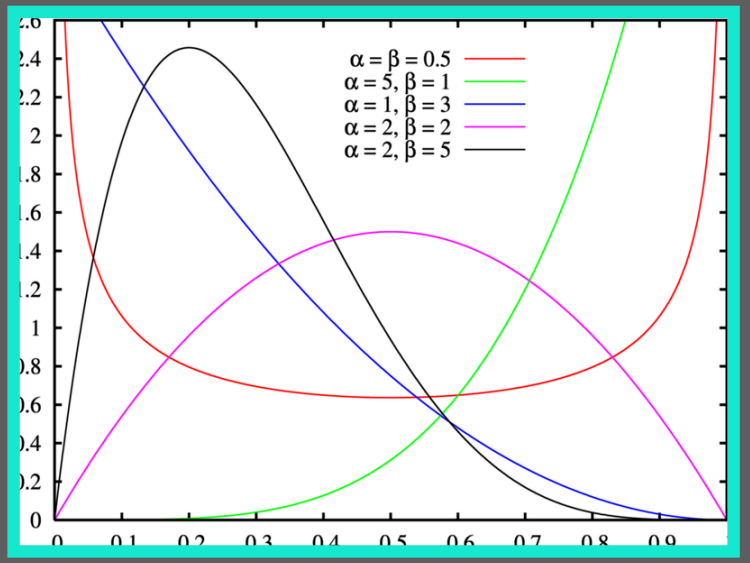
$$p_{\text{Beta}}(b \mid \alpha = k+1, \beta = N-k+1) \rightarrow$$

$$p_{\text{Binomial}}(k \mid b, N) = C_N^k b^k (1-b)^{N-k} = \frac{N!}{(N-k)!k!} b^k (1-b)^{N-k}$$

data (new): k; parameter: b



$0 < x < \infty,\ k > 0 \quad \bar{x} = k\theta$



$0 < x < 1,\ \alpha > 0,\ \beta > 0 \quad \bar{x} = \frac{\alpha}{\alpha+\beta}$

Bayesian Priors - II: Conjugate Priors

When the *posterior* and the *prior pdf* take the same functional form, then the prior is called a conjugate prior:

$$p(\vec{\theta} \mid D, I) \propto p(D \mid \vec{\theta}, I)\, p(\vec{\theta} \mid I)$$

New guess on parameter from new data

Old guess on parameter from previous data

Likelihood function

$$p_{\text{likelihood}}(x \mid \mu, \sigma) \sim \mathcal{N}(\mu, \sigma)$$

Experimental results Model parameter

For a Gaussian likelihood,

the conjugate prior $p(\vec{\theta} \mid I)$ *is also Gaussian:*

$$p_{\text{p}}(\mu, \sigma \mid x) \sim \mathcal{N}(\mu, \sigma)$$

Model parameter Experimental results

★ *Prior: in the case of previous experimental results*

★ *Posterior: in the case of new experimental results*

Bayesian Priors - II: Conjugate Priors

When the *posterior* and the *prior pdf* take the same functional form, then the prior is called a conjugate prior:

$$p(\vec{\theta} \mid D, I) \propto p(D \mid \vec{\theta}, I)\, p(\vec{\theta} \mid I)$$

New guess on parameter from new data

Old guess on parameter from previous data

Likelihood function

$$p_{\text{likelihood}}(m \mid \mu) \sim \frac{\mu^m \exp(-\mu)}{m!}$$

Experimental results Model parameter

For a Poisson likelihood,

the conjugate prior is a Gamma distribution:

$$p_{\text{p}}(\mu \mid k = m + 1, \theta = 1)$$

Model parameter Experimental results

★ *Prior: in the case of previous experimental results*

★ *Posterior: in the case of new experimental results*

Bayesian Priors - II: Conjugate Priors

When the *posterior* and the *prior pdf* take the same functional form, then the prior is called a conjugate prior:

$$p(\vec{\theta} \mid D, I) \propto p(D \mid \vec{\theta}, I)\, p(\vec{\theta} \mid I)$$

New guess on parameter from new data

Old guess on parameter from previous data

Likelihood function

$$p_{\text{likelihood}}(k \mid b, N) = C_N^k b^k (1-b)^{N-k} = \frac{N!}{(N-k)!k!} b^k (1-b)^{N-k}$$

Experimental results | Model parameter

For a binomial likelihood, the conjugate prior is a Beta distribution:

$$p_{\text{p}}(b \mid \alpha = k+1, \beta = N-k+1)$$

Model parameter | Experimental results

★ *Prior: in the case of previous experimental results*

★ *Posterior: in the case of new experimental results*

Hierarchical priors

Bayesian Priors - III: hyper-parameters and hyper-prior

Given a prior distribution, e.g.,:

$$p(\theta \mid \mu_0) = \frac{1}{\mu_0} \exp\left(\frac{-\theta}{\mu_0}\right)$$

$$p(\theta \mid \mu_0, V_0) = \frac{1}{\sqrt{2\pi V_0}} \exp\left(\frac{-(\theta - \mu_0)^2}{2V_0}\right)$$

What if the prior information parameters (e.g,. μ_0, V_0) are not explicitly given?

Recall previously, μ_0 and V_0 are ***prior information***, but are *not* model parameters! When μ_0 and V_0 are known, with the above-given prior distributions for model parameter θ, we are doing *standard Bayesian.*

$$p(\theta \mid D) = \frac{p(D \mid \theta, \mu_0, V_0)\, p(\theta \mid \mu_0, V_0)}{p(D)}$$

e.g. θ is the stellar mass of a galaxy (that we want to estimate through Bayesian inference), the prior μ_0, V_0 are the averaged stellar mass and its dispersion of a certain population.

Bayesian Priors - III: hyper-parameters and hyper-prior

Now prior information μ_0, V_0 may actually also depend on the dark matter halo mass M or evolve with redshift z, through further parameterized relationships:

e.g., $\lg \mu_0 = \alpha \lg M + \beta$ or $\mu_0 = \alpha z + \beta$,

where α, β are called "hyper-parameters", and they also have their "hyper-priors" $p(\alpha)$.

The inference now becomes *hierarchical Bayesian inference!*

Why do we need *hierarchical Bayesian inference?* — *I*

Sometimes the hyper parameters can be more interesting than the model parameters of individual systems!

Now we are no longer interested in θ, but interested in properties of hyper-parameters α, e.g., instead of stellar mass (as encoded by θ) of individual galaxies, we are interested in its averaging dependence on dark matter halo mass, or the average redshift evolution, which is encoded by parameter α, i.e., we want to obtain $p(\alpha|D)$. How do we do that?

Hierarchical Bayesian Inference

Lets consider the simplest case, where θ is model parameter and α is hyper-parameter.

We start with: $p(\theta|D) = \dfrac{p(D|\theta)p(\theta)}{p(D)}$ and $p(\theta) = \int p(\theta|\alpha)p(\alpha)\,\mathrm{d}\alpha$.

Priors of θ

Posterior pdf

i.e. $p(\theta|D) = \dfrac{p(D|\theta)\int p(\theta|\alpha)\,p(\alpha)\,\mathrm{d}\alpha}{p(D)}$.

Likelihood *Hyperpriors of α*

The inference will take us to obtain a best estimate of model parameter θ.

But if we want to obtain $p(\alpha|D)$. How do we obtain that?

Let's see first, $p(\alpha,\theta) = p(\theta|\alpha)p(\alpha) = p(\alpha|\theta)p(\theta)$.

Note: $p(\theta) = \int p(\alpha,\theta)\,\mathrm{d}\alpha$ and $p(\alpha) = \int p(\alpha,\theta)\,\mathrm{d}\theta$.

In the case of three parameters (notice the symmetry):

$$\begin{aligned} p(\alpha,\theta,D) &= p(D|\alpha,\theta)p(\theta|\alpha)p(\alpha) = p(\theta|\alpha,D)p(D|\alpha)p(\alpha) \\ &= p(D|\alpha,\theta)p(\alpha|\theta)p(\theta) = p(\alpha|\theta,D)p(D|\theta)p(\theta) \\ &= p(\theta|\alpha,D)p(\alpha|D)p(D) = p(\alpha|\theta,D)p(\theta|D)p(D) \end{aligned}$$

Hierarchical Bayesian Inference

Care for direct model parameters

Normally we have $p(\theta|D) = \dfrac{p(D|\theta)\int p(\theta|\alpha)\,p(\alpha)\,\mathrm{d}\alpha}{p(D)}$

Standard Bayesian

Conditional probability given D:

$$p(\alpha,\theta|D) = p(\theta|\alpha,D)p(\alpha|D) = \frac{p(\theta|\alpha,D)\;\;p(D|\alpha)p(\alpha)}{p(D)} = \frac{p(D|\alpha,\theta)p(\theta|\alpha)p(\alpha)}{p(D)}$$

Care for hyper-parameters

Marginalize over θ => $p(\alpha|D) = \int p(\alpha,\theta|D)\,\mathrm{d}\theta = \dfrac{\int p(D|\alpha,\theta)p(\theta|\alpha)p(\alpha)\,\mathrm{d}\theta}{p(D)}$

Hierarchical Bayesian

The SL2S Galaxy-scale Lens Sample. V. Dark Matter Halos and Stellar IMF of Massive Early-type Galaxies Out to Redshift 0.8

Show affiliations

https://arxiv.org/pdf/1410.1881.pdf

Sonnenfeld, Alessandro ; Treu, Tommaso ; Marshall, Philip J. ; Suyu, Sherry H. ; Gavazzi, Raphaël ; Auger, Matthew W. ; Nipoti, Carlo

A useful way to estimate scaling relations at population level.

Hierarchical Bayesian Inference

Why do we need *hierarchical Bayesian inference?* — II

Degeneracies can easily be present among parameters $\vec{\theta}$. However they are not easy to break due to limited number of observational constraints.

Without hierarchical inference, no priors of correlation are pre-imposed among model parameters $\vec{\theta}$. This, however, could result in *spurious* correlations among $\vec{\theta}$ in the final inferred joint posterior due to model degeneracy.

With hierarchical Bayesian, hyper-parameters $\vec{\alpha}$ are introduced to pre-describe/account for interconnections among model parameters $\vec{\theta}$. Such interconnections reflect the underlying relations among $\vec{\theta}$ at the population level. The existence of these relations can implicitly provide extra constraints and crucial knowledge to break degeneracies at a population level.

Dust Spectral Energy Distributions in the Era of Herschel and Planck: A Hierarchical Bayesian-fitting Technique

Show affiliations

https://arxiv.org/pdf/1203.0025.pdf

Kelly, Brandon C.; Shetty, Rahul; Stutz, Amelia M.; Kauffmann, Jens; Goodman, Alyssa A.; Launhardt, Ralf

A useful way to break degeneracy by imposing possible correlations among parameters at population level.

About Bayesian Priors

In many cases, Bayesian analysis —> Maximum likelihood estimates.

But in some other cases, the allowance of having
(a special type of) priors can significantly affect inferencing results.

This is in particular the case where ***selection effects*** (act as priors) matter.
Selection functions and sample truncation can cause ***selection bias:***

including *Malmquist bias*, where intrinsically brighter galaxies can
be seen to larger distances in a flux-limited galaxy survey;

and
Eddington bias, where more abundant intrinsically fainter galaxies
would, upon detection, enhance the number counts of intrinsically brighter
but fewer galaxies in the presence of non-negligible measurement errors.

3. Application of Bayesian inference:

Bayesian Parameter Estimate & Uncertainty

Bayesian Hypothesis Testing and Model Selection

Bayesian parameter estimates

We would like to infer the mean velocity μ^* and velocity dispersion σ^* of a stellar cluster. To do so, we have collected velocity measurements $\{x_i\}$ of N individual stars in that stellar cluster.

Let's look at this simplest set up as an example case and build up our complexity level on the error behavior gradually …

$$p(\vec{\theta} \mid D) = \frac{p(D \mid \vec{\theta}, I)\, p(\vec{\theta} \mid I)}{p(D)}$$

Case I: the measurements of a property $\{x_i\}$ with ***known*** heteroscedastic error $\{\sigma_i\}$ (measurement error), we seek for the posterior pdf $p(\mu \mid \{x_i\}, \{\sigma_i\})$.

Case II: when intrinsic scatter σ, which is much larger than individual measurement error, is ***unknown***, seek for a 2d posterior $p(\mu, \sigma \mid \{x_i\}, I)$.

Case III: measurement errors $\{e_i\}$ are also known and non-negligible, we seek for a 2d posterior pdf $p(\mu, \sigma \mid \{x_i\}, \{e_i\})$.

Bayesian parameter estimates

Case I: the measurements of a property $\{x_i\}$ follow a Gaussian distribution and have ***known*** heteroscedastic ***errors*** $\{\sigma_i\}$, i.e., we seek for the posterior pdf $p(\mu \,|\, \{x_i\}, \{\sigma_i\})$.

$$p(\vec{\theta} \,|\, D) = \frac{p(D \,|\, \vec{\theta}, I)\, p(\vec{\theta} \,|\, I)}{p(D)}$$

Data likelihood $$p(\{x_i\} \,|\, \mu, I) = \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma_i^2}\right)$$

Prior of μ $$p(\mu \,|\, I) = C, \text{ for } \mu_{\min} \leqslant \mu \leqslant \mu_{\max}$$
$$\text{where } C = (\mu_{\max} - \mu_{\min})^{-1}$$

Logarithm posterior $$L_p = \ln[p(\mu \,|\, \{x_i\}, \{\sigma_i\}, I)] = \text{constant} - \Sigma_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma_i^2}$$

Bayesian parameter estimates

Case I: ***known*** $\{\sigma_i\}$, seek for the posterior pdf $p(\mu \mid \{x_i\}, \{\sigma_i\})$

To apply ***maximum a posteriori (MAP)*** estimate:

Requesting $\left(\mathrm{d}L_p/\mathrm{d}\mu\right)|_{(\mu=\mu_0)} = 0$

$$\mu_0 = \frac{\Sigma_{i=1}^{N} \omega_i x_i}{\Sigma_{i=1}^{N} \omega_i} \qquad \omega_i = \sigma_i^{-2}$$

Identical to the *MLE* result
In classical inference

$$\sigma_\mu = \left(- \frac{\mathrm{d}^2 L_p}{\mathrm{d}\mu^2} \Big|_{\mu=\mu_0} \right)^{-1/2} = \left(\Sigma_{i=1}^{N} \omega_i \right)^{-1/2} = \left(\Sigma_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1/2}$$

It can be shown that when $\{\sigma_i\}$ are know, higher-order terms in the Taylor expansion of L_p vanish after the 2^{nd} , therefore the posterior *pdf* is a Gaussian of $\mu \sim \mathcal{N}(\mu_0, \sigma_\mu)$.

Bayesian parameter estimates

Case II: when scatter/error σ is ***unknown*** (this can either be an intrinsic scatter or some dominant homoscedastic measurement error), to be determined also from data, i.e., we seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$.

Data likelihood $$p(\{x_i\} \,|\, \mu, \sigma, I) = \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right)$$

Prior of μ, σ: $p(\mu, \sigma \,|\, I) \propto C$, for $\mu_{\rm min} \leqslant \mu \leqslant \mu_{\rm max}$ and . $\sigma_{\rm min} \leqslant \sigma \leqslant \sigma_{\rm max}$

Option 1 flat prior for μ and σ

Posterior: $$p(\{x_i\} \,|\, \mu, \sigma, I)\, p(\mu, \sigma \,|\, I) = C \frac{1}{\sigma^N} \Pi_{i=1}^{N} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right)$$

$$C = (2\pi)^{-N/2} (\mu_{\rm max} - \mu_{\rm min})^{-1} (\sigma_{\rm max} - \sigma_{\rm min})^{-1}$$

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, I)] = \text{constant} - N \ln \sigma - \Sigma_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - N \ln \sigma - \Sigma_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Note that $\Sigma_{i=1}^{N}(x_i - \mu)^2 = N(\bar{x} - \mu)^2 + NV$

where $\bar{x} = N^{-1}\Sigma_{i=1}^{N} x_i$ (sample mean),

and $V \equiv N^{-1}\Sigma_{i=1}^{N}(x_i - \bar{x})^2 = (N-1)s^2/N$

where s^2 is the sample variance.

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - N \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

In this case, $(N, \bar{x}, V)$ fully capture the information content of the posterior!!

sample statistics: summary statistics about the measurement

The position (μ_0, σ_0) of the *maximum* joint posterior is obtained by requesting $\left(\mathrm{d}L_p/\mathrm{d}\mu\right)|_{(\mu=\mu_0)} = 0$ and $\left(\mathrm{d}L_p/\mathrm{d}\sigma\right)|_{(\sigma=\sigma_0)} = 0$:

$\mu_0 = \bar{x}$ [sample mean] and $\sigma_0^2 = V$ [nearly sample variance]

Alternatively

Bayesian parameter estimates

Case II: when scatter/error σ is ***unknown*** (this can either be an intrinsic scatter or some dominant homoscedastic measurement error), to be determined also from data, i.e., we seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$.

Data likelihood $$p(\{x_i\} \,|\, \mu, \sigma, I) = \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

Prior of μ, σ: $$p(\mu, \sigma \,|\, I) \propto 1/\sigma, \text{ for } \mu_{\min} \leqslant \mu \leqslant \mu_{\max} \text{ and . } \sigma_{\min} \leqslant \sigma \leqslant \sigma_{\max}$$

flat prior for μ and scale-free prior for σ

Posterior: $$p(\{x_i\} \,|\, \mu, \sigma, I)\, p(\mu, \sigma \,|\, I) = C \frac{1}{\sigma^{N+1}} \Pi_{i=1}^{N} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

$$C = (2\pi)^{-N/2} (\mu_{\max} - \mu_{\min})^{-1} \left[\ln\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)\right]^{-1}$$

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, I)] = \text{constant} - (N+1)\ln\sigma - \Sigma_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Alternatively

Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - (N + 1)\ln\sigma - \Sigma_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Note that $\Sigma_{i=1}^{N}(x_i - \mu)^2 = N(\bar{x} - \mu)^2 + NV$

where $\bar{x} = N^{-1}\Sigma_{i=1}^{N} x_i$ (sample mean),

and $V \equiv N^{-1}\Sigma_{i=1}^{N}(x_i - \bar{x})^2 = (N-1)s^2/N$

where s^2 is the sample variance.

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - (N + 1)\ln\sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

In this case, $(N, \bar{x}, V)$ fully capture the information content of the posterior!!

sample statistics

The position (μ_0, σ_0) of the *maximum* joint posterior is obtained by requesting $\left(\mathrm{d}L_p/\mathrm{d}\mu\right)|_{(\mu=\mu_0)} = 0$ and $\left(\mathrm{d}L_p/\mathrm{d}\sigma\right)|_{(\sigma=\sigma_0)} = 0$:

$\mu_0 = \bar{x}$ [sample mean] and $\sigma_0^2 = VN/(N+1)$ [nearly sample variance]

Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$

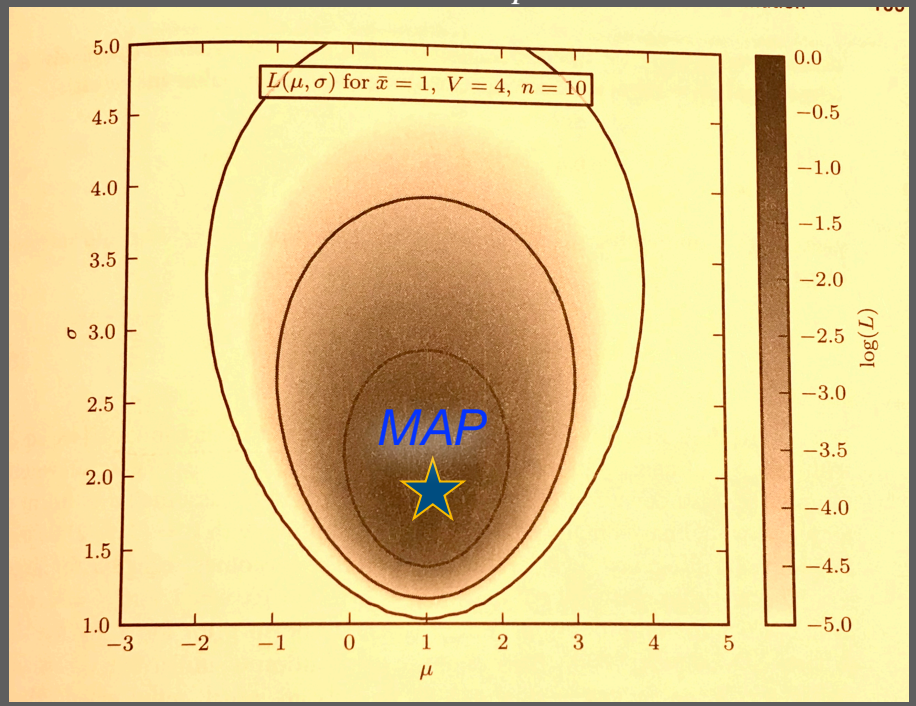
With flat prior for μ and σ:

$$L_{1p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - N \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

With flat prior for μ and scale-free prior for σ:

$$L_{2p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - (N+1) \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

2d logarithm posterior $L_{2p}(\mu, \sigma)$



MAP of the joint posterior L_{1p}:

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{V})$

MAP of the joint posterior L_{2p}:

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{VN/(N+1)})$

where $V \equiv (N-1)s^2/N$ and

$$s^2 \equiv \frac{\Sigma_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \;:\; \text{sample variance}$$

What are the posterior distributions for μ and σ after marginalizing over the other parameter? Are they still Gaussian?

Review

ent's *t* distribution

.S. Gosset

$$p(t\,|\,k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

egree of freedom

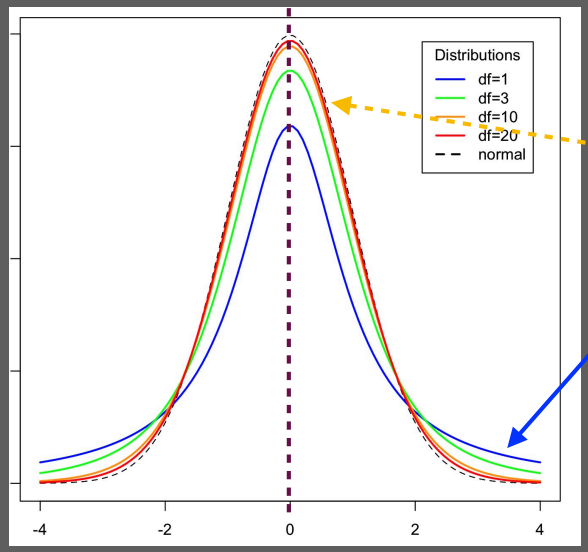
Eh, I need a distribution for small samples, and I need a Guinness when doing statistics...

$p(t) = \pi^{-1}(1+t^2)^{-1}$ becomes

y distributions with $\mu = 0,\ \gamma = 1$

: $x \sim \mathcal{N}(0,1)$ - becomes a

standard normal distribution.



Heavy tails, symmetric bell-shape about 0

$k > 1:\ \bar{t} = 0$

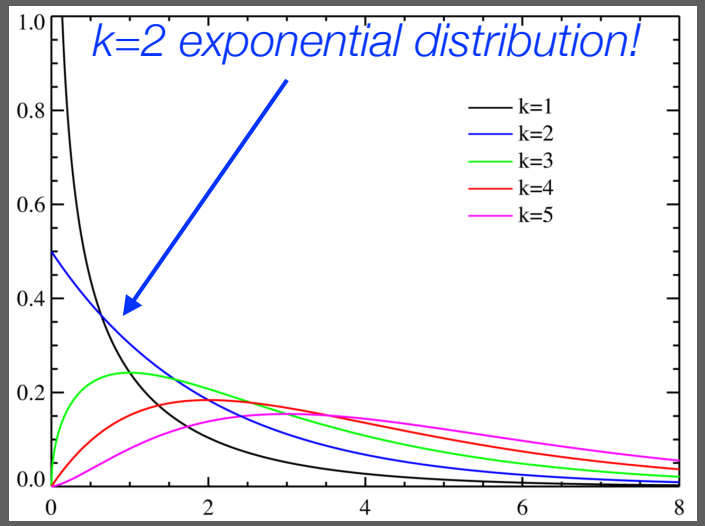
$k > 2:\ \sigma_t = \sqrt{k/(k-2)}$, but $k \leqslant 2$, no variance σ_t

$k > 3:\ \Sigma = 0$, but $k \leqslant 3$, no skewness Σ

$k > 4:\ K = 6/(k-4)$, but $k \leqslant 4$, no kurtosis K

Review

χ^2 distribution



Given a random variable $x \sim \mathcal{N}(\mu, \sigma)$,

If we define $z \equiv \dfrac{x-\mu}{\sigma}$, then $z \sim \mathcal{N}(0,1)$.

Now draw random sets each time with size of n, and each time calculate $Q \equiv \Sigma_{i=1}^{n} z_i^2 \ (>0)$, then Q follows a χ^2 distribution with $k = n$ degree of freedom:

$$p(Q\,|\,k) \equiv \chi^2(Q\,|\,k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{k/2-1} \exp(-Q/2)$$

Not depend on actual value of μ, σ!

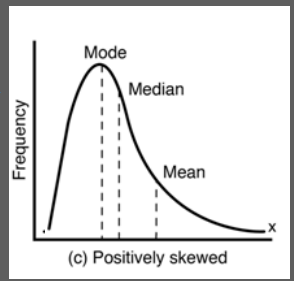
Gamma function: $\Gamma(k/2) = \dfrac{(k-2)!!\sqrt{\pi}}{2^{(k-1)/2}}$ — a positive half-integer

1. If $Q_j \ (j = 1, 2, \ldots M)$ are a series of χ^2 distributions with $n_1, n_2, \ldots n_M$ degree of freedom, and are mutually independent, then $\Sigma_{j=1}^{M} Q_j \sim \chi^2(n_1 + n_2 + \ldots + n_M)$.
   i.e., Q_j follows a χ^2 with $k = n_j$ and the sum $\Sigma_{j=1}^{M} Q_j$ also follows a χ^2 with $k = \Sigma_{j=1}^{M} n_j$
   — Combining independent datasets to constrain the same underlying model!
2. $\chi^2_{\rm dof}(Q_{\rm red}\,|\,k) \equiv P(Q/k\,|\,k)$ is called ***reduced χ^2 distribution, where*** $Q_{\rm red} \equiv Q/k$:

   $\overline{Q_{\rm red}} = 1, \sigma_{Q_{\rm red}} = \sqrt{2/k}, \Sigma = \sqrt{8/k}, K = 12/k,$

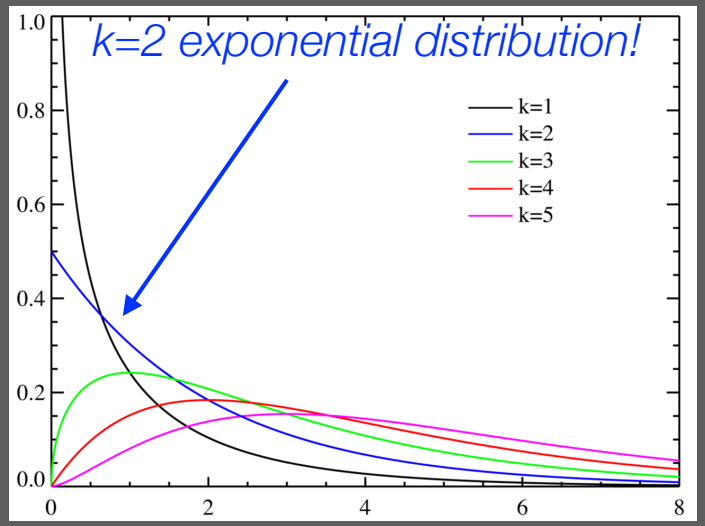
   as $k \Uparrow$, $\chi^2_{\rm dof}(Q_{\rm red}\,|\,k) \ \rightarrow \mathcal{N}(1, \sqrt{2/k})$.

used in maximum likelihood method



Review

χ^2 distribution



Given a random variable $x \sim \mathcal{N}(\mu, \sigma)$,

If we define $z \equiv \dfrac{x-\mu}{\sigma}$, then $z \sim \mathcal{N}(0,1)$.

Now draw random sets each time with size of n, and each time calculate $Q \equiv \Sigma_{i=1}^{n} z_i^2 \ (>0)$, then Q follows a χ^2 distribution with $k = n$ degree of freedom:

$$p(Q\,|\,k) \equiv \chi^2(Q\,|\,k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{k/2-1} \exp(-Q/2)$$

Not depend on actual value of μ, σ!

Gamma function: $\Gamma(k/2) = \dfrac{(k-2)!!\sqrt{\pi}}{2^{(k-1)/2}}$ (a positive half-integer)

2. $\chi^2_{\rm dof}(Q_{\rm red}\,|\,k) \equiv P(Q/k\,|\,k)$ is called *reduced χ^2 distribution, where* $Q_{\rm red} \equiv Q/k$:

$\overline{Q_{\rm red}} = 1, \sigma_{Q_{\rm red}} = \sqrt{2/k}, \Sigma = \sqrt{8/k}, K = 12/k,$ *used in maximum likelihood method*

as $k \Uparrow$, $\chi^2_{\rm dof}(Q_{\rm red}\,|\,k) \ \rightarrow \mathcal{N}(1, \sqrt{2/k})$.

Note: *χ^2 analysis is sensitive to outliers, see future lecture to deal with non-Gaussian errors.*

3. In the context of model fitting to measurements with *Gaussian errors σ*,

$$Q_{\rm red} \equiv Q/k = \frac{\Sigma_{i=1}^{n} z_i^2}{N_{\rm point} - N_{\rm param}}$$

for $k = 200$, $\Delta P(Q_{\rm red} \in [0.9 - 1.1]) = 68\,\%$

for $k = 20{,}000$, $\Delta P(Q_{\rm red} \in [0.99 - 1.01]) = 68\,\%$

If a model makes "good" prediction of μ, then $Q_{\rm red} \sim 1$ in particular when dataset is large.

Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$

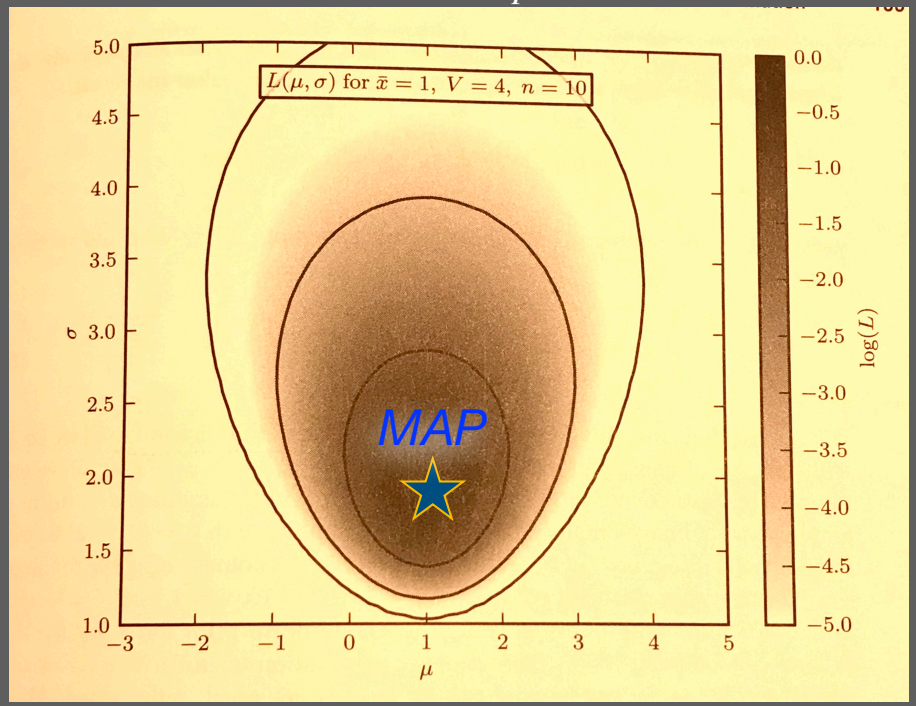
With flat prior for μ and σ:

$$L_{1p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - N \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

With flat prior for μ and scale-free prior for σ:

$$L_{2p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - (N+1) \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

2d logarithm posterior $L_{2p}(\mu, \sigma)$



MAP of the joint posterior L_{1p}:

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{V})$

MAP of the joint posterior L_{2p}:

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{VN/(N+1)})$

where $V \equiv (N-1)s^2/N$ and

$$s^2 \equiv \frac{\Sigma_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \;:\; \text{sample variance}$$

What are the posterior distributions for μ and σ after marginalizing over the other parameter? Are they still Gaussian?

Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$

With flat prior for μ and σ:

$$L_{1p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - N \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

With flat prior for μ and scale-free prior for σ:

$$L_{2p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - (N+1) \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

★ To obtain the posterior probability $p(\mu)$, marginalizing over all possible σ:

$$p(\mu \,|\, \{x_i\}, I) = \int_0^\infty p(\mu, \sigma \,|\, \{x_i\}, I)\, \mathrm{d}\sigma \propto \left[1 + \frac{(\mu - \bar{x})^2}{V}\right]^{-N/2}$$

This posterior corresponds to a ***Student's t distribution*** with

$N - 1$ degrees of freedom for $t = (\mu - \bar{x})/(s/\sqrt{N})$.

[note: only at large N limit —> normal distribution]

Note: maximizing marginalized $p(\mu)$ also yields MAP estimate $\mu_0 = \bar{x}$ [sample mean]

Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$

With flat prior for μ and σ:

$$L_{1p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - N \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

★ To obtain the posterior probability $p(\sigma)$, marginalizing over all μ:

$$p(\sigma \,|\, \{x_i\}, I) = \int_0^\infty p(\mu, \sigma \,|\, \{x_i\}, I)\, \mathrm{d}\mu \propto \frac{1}{\sigma^{N-1}} \exp\left(\frac{-NV}{2\sigma^2}\right)$$

Corresponding to a χ^2 ***distribution*** with $N + 1$ degrees of freedom for $Q = NV/\sigma^2$.

maximizing (marginalized) $p(\sigma)$ yields $\sigma_0^2 = s^2$ [sample variance!]

With flat prior for μ and scale-free prior for σ:

$$L_{2p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - (N+1) \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

★ To obtain the posterior probability $p(\sigma)$, marginalizing over all μ:

$$p(\sigma \,|\, \{x_i\}, I) = \int_0^\infty p(\mu, \sigma \,|\, \{x_i\}, I)\, \mathrm{d}\mu \propto \frac{1}{\sigma^{N}} \exp\left(\frac{-NV}{2\sigma^2}\right)$$

Corresponding to a χ^2 ***distribution*** with $N + 2$ degrees of freedom for $Q = NV/\sigma^2$.

maximizing (marginalized) $p(\sigma)$ yields $\sigma_0^2 = V$ [nearly sample variance]

Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, I)$

With flat prior for μ and σ:

$$L_{1p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - N \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

MAP of the joint posterior L_{1p}*:*

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{V})$

MAP of the marginalized posteriors:

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{s})$

With flat prior for μ and scale-free prior for σ:

$$L_{2p} = \ln[p(\mu, \sigma \,|\, \{x_i\}, I) = \text{constant} - (N+1) \ln \sigma - \frac{N}{2\sigma^2}\left((\mu - \bar{x})^2 + V\right)$$

MAP of the joint posterior L_{2p}*:*

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{VN/(N+1)})$

MAP of the marginalized posteriors:

★ $(\mu_0, \sigma_0) = (\bar{x}, \sqrt{V})$

The marginalized $p(\mu)$ follows a ***Student's*** t ***distribution***

The marginalized $p(\sigma)$ follows a χ^2 ***distribution***

How different is the confidence intervals between the Bayesian inference and classical inference?

To find best model parameters:
Maximum likelihood estimate (MLE)

To quantify parameter uncertainties:
— Confidence regions:

1. Through Fisher Information matrix
2. Bootstrap or Jackknife (resampling)

Classical/frequentist perspective

Strongly based on central limit theorem which states that when the sample size is large, the distribution converges to Gaussian!

Central Limit Theorem

The theorem states that *when the sample size n is large*, regardless of the shape of the underlying population's distribution (as long as the population has mean μ_X and standard deviation σ_X), *the sample mean $\overline{X}$ and standard deviation S (for a given realization) follow Gaussian distributions, the mean of which converge to the population statistics, with uncertainties decreasing as $\sigma_{\overline{x},\,S} \sim \sigma_X\, n^{-1/2}$.*

Specifically, when $n \to \infty$, $\overline{X} \sim \mathcal{N}(E(\overline{X}), \sigma_{\overline{X}})$, where $E(\overline{X}) = \mu_X$, and $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}}$.

The latter $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}}$ can be estimated by:

$$\sigma'_{\overline{X}} \equiv \frac{S}{\sqrt{n}}$$ *(standard error of the mean)*

SEM: $\sigma'_{\overline{X}} = \dfrac{S}{\sqrt{n}}$: $\overline{X} \pm 2\sigma'_{\overline{X}}$ [95%]

Similarly, when $n \to \infty$, $S \sim \mathcal{N}(E(S), \sigma_S)$, where $E(S) = \sigma_X$, $\sigma_S = \dfrac{\sigma_X}{\sqrt{2(n-1)}}$.

The latter can be estimated by $\sigma'_S \equiv \dfrac{S}{\sqrt{2(n-1)}}$ — *error of sample standard deviation*.

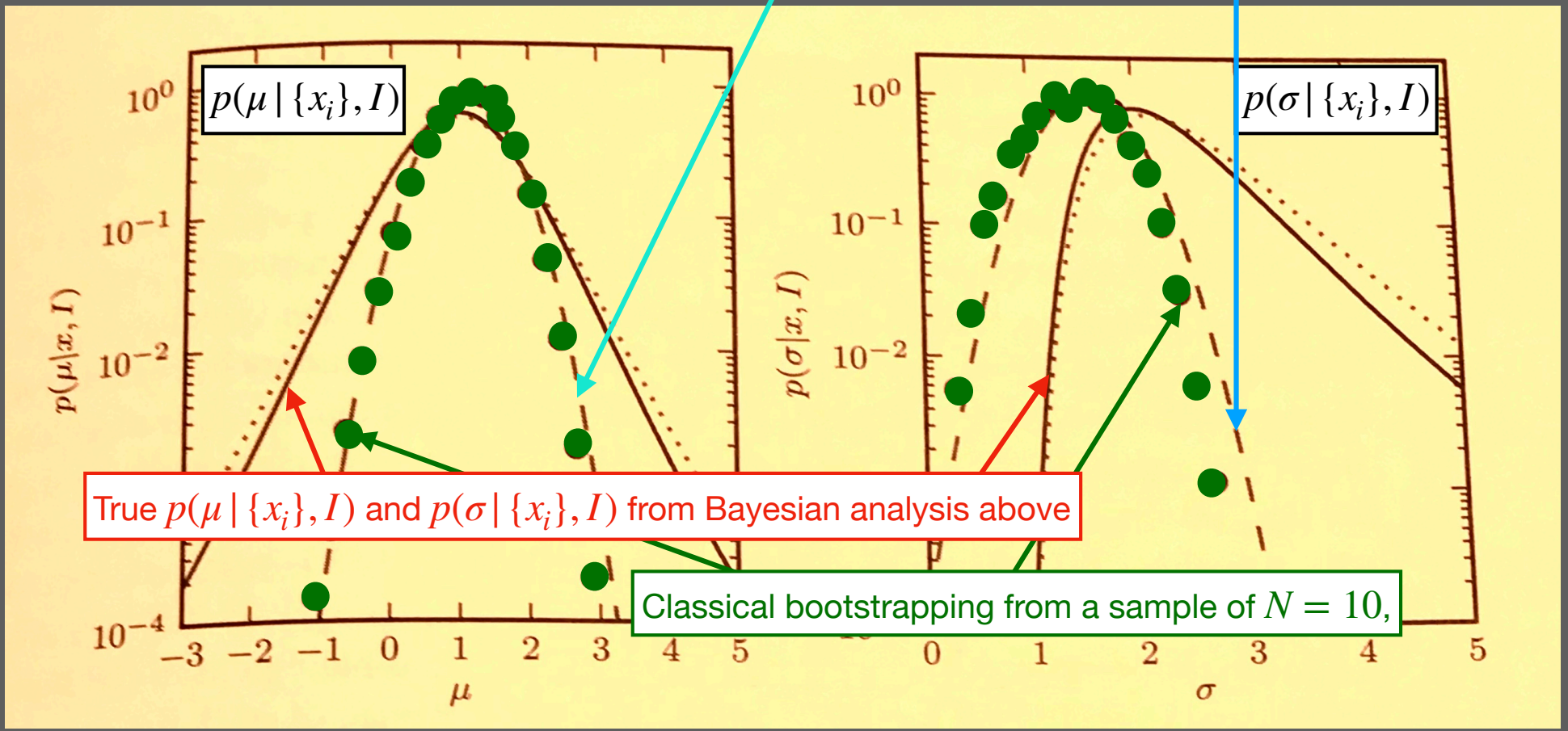
Note: σ_X is the standard deviation of the population distribution of random variable X, it is a physical property - intrinsic scatter. While $\sigma_{\overline{X}}$ is the standard deviation of the measured sample mean $\overline{X}$ over many realizations/draws of the underlying population, which goes to zero as $n \to \infty$.

In simple words, as the sample size increases, the sample mean and standard deviation all converge to the population statistics (*Bernoulli's "Law of large number"*).

Bayesian parameter estimates When the sample size is small …

When N is large, all three estimates converge to Gaussian statistics.

When N is small (i.e., $N < 100$ for most applications), the classical estimates based on ***CLT*** (i.e., assuming Gaussian statistics or resampling-based methods) would fail, and the derived confidence intervals strongly differ from the true ones!



Bayesian parameter estimates

Case II: ***unknown*** σ, seek for a 2d posterior pdf $p(\mu, \sigma | \{x_i\}, I)$

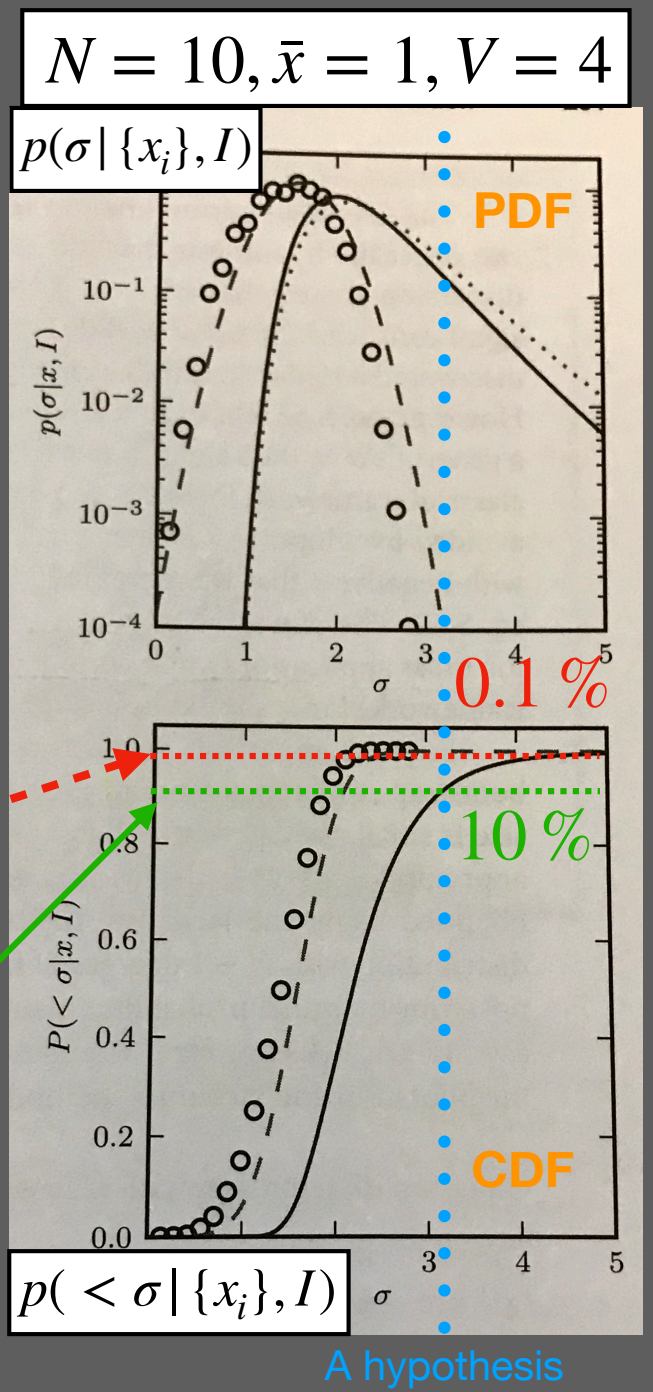
When N is small, such differences can result in quantitatively different conclusion in hypothesis testing!

Consider this following question:

Can you rule out a model with $\sigma > 3$ at a 5% significance level?

A classical inference under *Gaussian* approximation suggests that such a probability is 0.1%.

But a proper *Bayesian* analysis will tell us one ***cannot*** reject the hypothesis at 5% significance level!



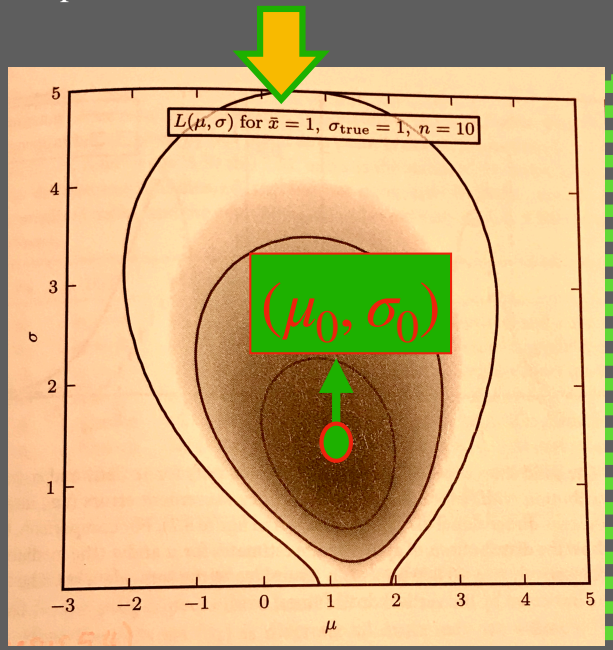
Bayesian parameter estimates

Case III: measurement error $\{e_i\}$ is known but intrinsic σ is ***unknown***, i.e., we seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, \{e_i\})$.

Data likelihood $$p(\{x_i\}, \{e_i\} \,|\, \mu, \sigma, I) = \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi}(\sigma^2 + e_i^2)^{1/2}} \exp\left(\frac{-(x_i - \mu)^2}{2(\sigma^2 + e_i^2)}\right)$$

Prior of μ, σ: $$p(\mu, \sigma \,|\, I) = \text{const, for } \mu_{\min} \leqslant \mu \leqslant \mu_{\max} \text{ and } \sigma_{\min} \leqslant \sigma \leqslant \sigma_{\max}$$

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, \{e_i\}, I) = \text{constant} - \frac{1}{2}\Sigma_{i=1}^{N}\left(\ln(\sigma^2 + e_i^2) + \frac{(x_i - \mu)^2}{(\sigma^2 + e_i^2)}\right)$$



Requesting $\left(\mathrm{d}L_p/\mathrm{d}\mu\right)|_{(\mu=\mu_0)} = 0$

$$\mu_0 = \frac{\Sigma_{i=1}^{N}\omega_i x_i}{\Sigma_{i=1}^{N}\omega_i} \quad \omega_i = \frac{1}{(\sigma_0^2 + e_i^2)}$$

Requesting $\left(\mathrm{d}L_p/\mathrm{d}\sigma\right)|_{(\sigma=\sigma_0)} = 0$

$$\Sigma_{i=1}^{N}\frac{1}{\sigma_0^2 + e_i^2} = \Sigma_{i=1}^{N}\frac{(x_i - \mu_0)^2}{(\sigma_0^2 + e_i^2)^2}$$

One can go for MAP on joint posterior can be solved iteratively!

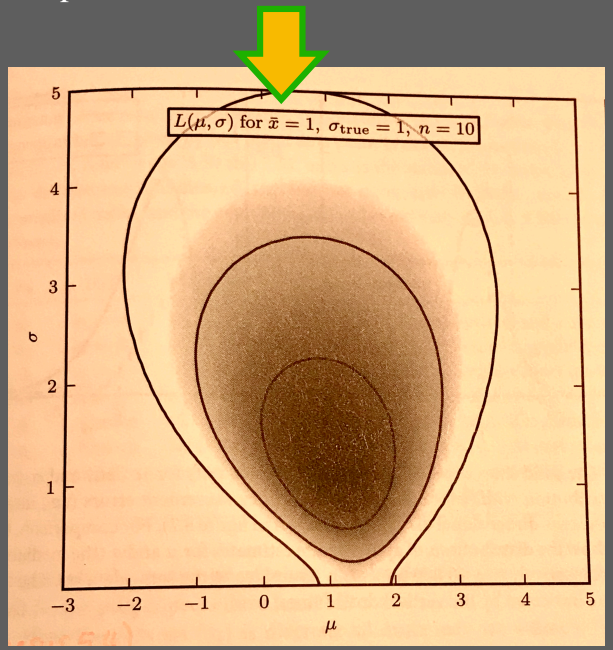
Bayesian parameter estimates

Case III: measurement error $\{e_i\}$ is known but intrinsic σ is ***unknown***, i.e., we seek for a 2d posterior pdf $p(\mu, \sigma \,|\, \{x_i\}, \{e_i\})$.

Data likelihood $$p(\{x_i\}, \{e_i\} \,|\, \mu, \sigma, I) = \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi}(\sigma^2 + e_i^2)^{1/2}} \exp\left(\frac{-(x_i - \mu)^2}{2(\sigma^2 + e_i^2)}\right)$$

Prior of μ, σ: $$p(\mu, \sigma \,|\, I) = \text{const, for } \mu_{\min} \leqslant \mu \leqslant \mu_{\max} \text{ and } \sigma_{\min} \leqslant \sigma \leqslant \sigma_{\max}$$

$$L_p = \ln[p(\mu, \sigma \,|\, \{x_i\}, \{e_i\}, I) = \text{constant} - \frac{1}{2}\Sigma_{i=1}^{N}\left(\ln(\sigma^2 + e_i^2) + \frac{(x_i - \mu)^2}{(\sigma^2 + e_i^2)}\right)$$



Or estimates based on marginalized posteriors. Note: in the case of asymmetric distribution, ***median*** *and interquartile range* σ_G *can be used as alternatives to describe the posterior distribution.*

interquartile range $\text{IQR} \equiv q_{75} - q_{25}$

$\sigma_{\text{G}} \equiv 0.7413\,\text{IQR}$

with $\sigma_G = \sigma$ for a Gaussian.

Bayesian parameter estimates

Upgrading our game!

With mixture model: identify and throw out “outliers”!

Same as before, we have a set of N measurements $\{x_i, \sigma_i\}$ for property μ (unknown). For most measurements, the errors follow Gaussian $\mathcal{N}(\mu, \sigma)$.
But now there are data points that do not follow $\mathcal{N}(\mu, \sigma)$, we call them "outliers" and we are uncertain regarding which data points are!

We introduce $\{g_i \in [0, 1]\}$ describing the probability for the $i^{\rm th}$ measurement to be "good", i.e., following a Gaussian error behavior; while $\{1 - g_i\}$ describing the probability of "bad", i.e., following the error function for outliers. $\{g_i\}$ are *unknown* (as we do *NOT* know which ones are outliers!) and shall be constrained (because we want to throw them away!).

$g_i = 1$, surely OK points!

$g_i = 0$, surely outliers!

For simplicity (without loss of generality), assuming that good measurements follow $\mathcal{N}(\mu, \sigma_i)$ and the outliers follow $\mathcal{N}(\mu, \sigma^*)$ (no bias), where some systematic error σ^* is significantly larger than typical σ_i. The data likelihood for each data point:

$$p(x_i, \sigma_i \,|\, \mu, g_i, I) = g_i \mathcal{N}(\mu, \sigma_i) + (1 - g_i)\, \mathcal{N}(\mu, \sigma^*)$$

The *prior* for the two unknown parameter sets reads:

$$p(\mu, \{g_i\} \,|\, I) = p(\mu \,|\, I)\, p(\{g_i\} \,|\, I)$$ Assume independence!

Then 2d posterior for $(\mu, \{g_i\})$ writes:

$$p(\mu, \{g_i\} \,|\, \{x_i\}, \{\sigma_i\}, I) \propto \underbrace{\Pi_{i=1}^{N}[g_i N(\mu, \sigma_i) + (1 - g_i) N(\mu, \sigma^*)]}_{\text{Likelihood}} \underbrace{p(\mu \,|\, I)}_{\text{Prior of } \mu} \underbrace{p(\{g_i\} \,|\, I)}_{\text{Prior of } \{g_i\}}$$

Marginalizing over all g_i to obtain the marginalized posterior probability for μ:

note: under uniform assumption on g_i this is effectively replacing each g_i with 1/2 in integration)

$$p(\mu \,|\, \{x_i\}, \{\sigma_i\}) \propto \int p(\mu, \{g_i\} \,|\, \{x_i\}, \{\sigma_i\}, I)\, \mathrm{d}^N g_i \propto \Pi_{i=1}^{N}[N(\mu, \sigma_i) + N(\mu, \sigma^*)]$$

We can also obtain estimate for g_j, answering how likely data point j is an outlier:

1. marginalizing over *all* $g_{i \neq j}$ to obtain a joint posterior probability for $\{\mu, g_j\}$:

$$p(\mu, g_j \,|\, \{x_i\}, \{\sigma_i\}) = [g_j N(\mu, \sigma_j) + (1 - g_j) N(\mu, \sigma^*)]\, \Pi_{i \neq j}^{N}[N(\mu, \sigma_i) + N(\mu, \sigma^*)]$$

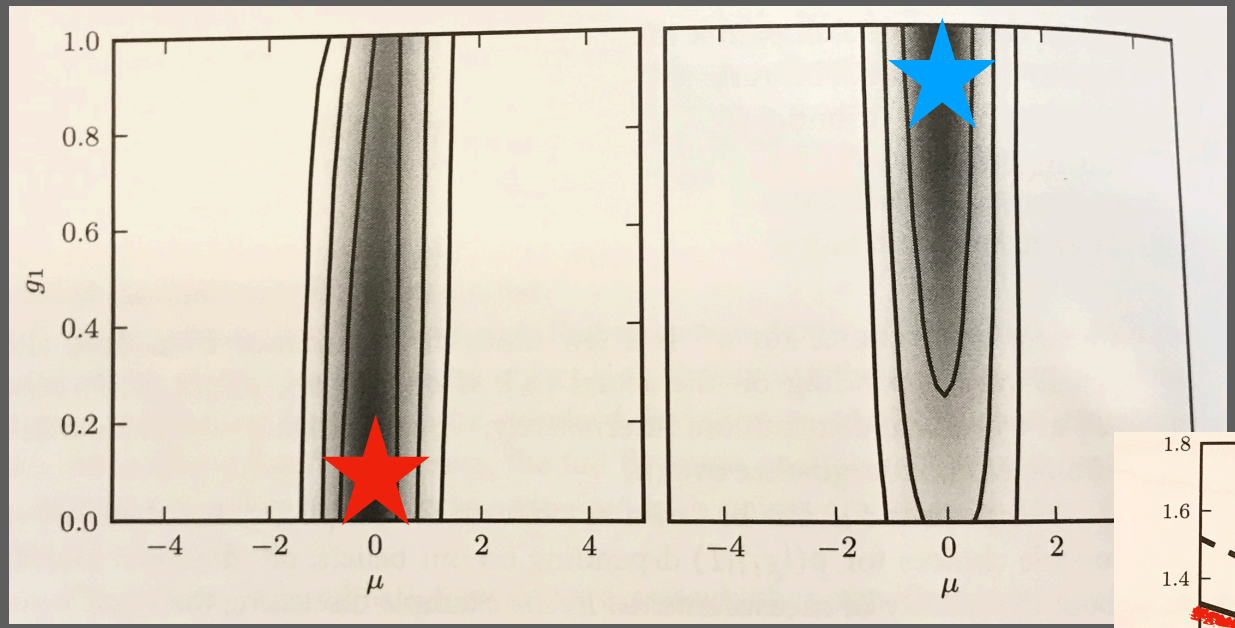
2. then marginalizing over μ to obtain estimate for g_j:

$$p(g_j \,|\, \{x_i\}, \{\sigma_i\}) = \int p(\mu, g_j \,|\, \{x_i\}, \{\sigma_i\})\, \mathrm{d}\mu$$

The joint posterior probability for $\{\mu, g_j\}$:

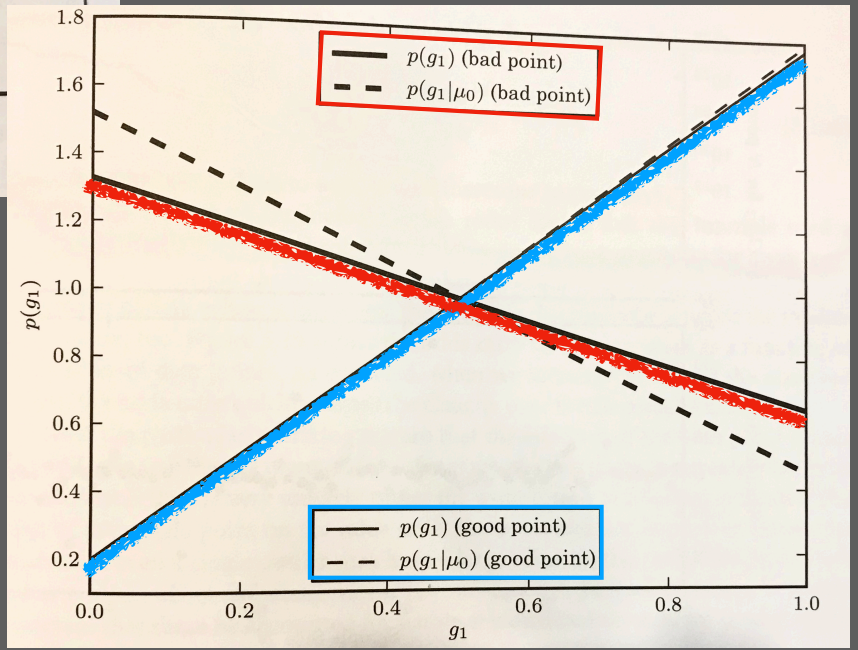
$$p(\mu, g_j \,|\, \{x_i\}, \{\sigma_i\}) = [g_j N(\mu, \sigma_j) + (1 - g_j) N(\mu, \sigma^*)] \, \Pi_{i \neq j}^{N} [N(\mu, \sigma_i) + N(\mu, \sigma^*)]$$

For figures below: a sample of 10 points, with 8 "good" points drawn from $\mathcal{N}(0, 1)$ and 2 "bad" (outliers) points drawn from $\mathcal{N}(0, 3)$.



$g_i = 1$, surely OK points!

$g_i = 0$, surely outliers!



Case 1: Point 1 is designed to be "bad", we get g_1 peaks at 0

Case 2: Point 1 is "good", g_1 peaks at 1

After marginalizing over μ to obtain estimate for g_j:

$$p(g_j \,|\, \{x_i\}, \{\sigma_i\}) = \int p(\mu, g_j \,|\, \{x_i\}, \{\sigma_i\}) \, \mathrm{d}\mu$$

Clarification

A Gaussian data likelihood

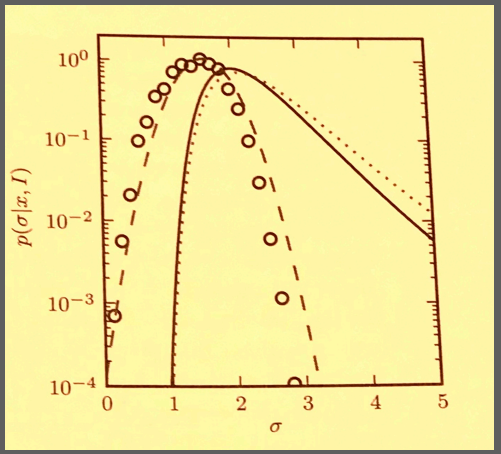
$$p(\{x_i, \sigma_i\} \mid \vec{\theta}, I) = \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(x_i - \mu_i(\vec{\theta}))^2}{2\sigma_i^2}\right)$$

Let us NOT get confused among 3 things below:

(1) the measurements $\{x_i\}$ follow a Gaussian distribution — this is an error behavior.

(2) The model $\mu(\vec{\theta})$ which generate/predict the theoretical value $\mu_i(\vec{\theta})$ corresponding to the i^{th} data point, can be in principle extremely complicated, but the error behavior may still be a simple Gaussian!

(3) The (marginalized) posterior pdf, e.g., $p(\theta \mid \{x_i\}, \{\sigma_i\})$ is *NOT* necessarily Gaussian at all!!



No matter how complicated your model is, end of the day, the likelihood function p is all simple and basic! It is based on the error behavior!

Typical error behaviors for likelihood and the conjugate prior

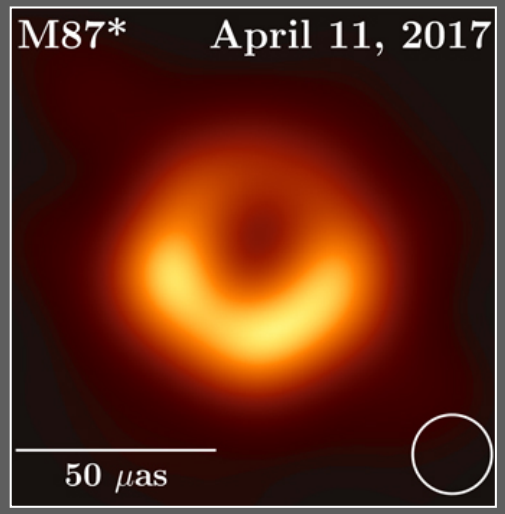
1. Gaussian (large number statistics) — Gaussian distribution
2. Poisson (small number statistics) — Gamma distribution
3. Binomial (success rate, “yes” or “no”) — Beta distribution

3. Application of Bayesian inference:

Bayesian Parameter Estimate & Uncertainty

Bayesian Hypothesis Testing and Model Selection

Bayesian Hypothesis Testing



Consider another example:

Theorists predict a 10% upper limit for the fraction of galaxies that host ***active*** supermassive black holes in local Universe.

Observers studied 10 galaxies and found evidence of ***active*** black holes in 4 of them.

Can we reject the theory at a significant level of $\alpha = 1\,\%$?

In this case, we consider a model which predict the ***active*** black-hole hosting "success" rate to be b. The *likelihood* that we find k $(k = 4)$ successful outcomes out of a total N $(N = 10)$ measurements under such a model is given by a ***binomial*** distribution:

$$p(k \,|\, b, N) = C b^k (1 - b)^{N-k}.$$

Assuming a flat prior for $b \in [0, 1]$, the ***posterior*** probability for b is given by:

$$p(b \,|\, k, N) = C b^k (1 - b)^{N-k}$$

* alternatively we can also use the Beta distribution as conjugate prior…

The null hypothesis is that the data is compatible with a black-hole hosting model which predicts $b \leqslant 0.1$. — *to be tested!*

Bayesian Hypothesis Testing

We ask for a given observation of $N = 10$ *measurements with* $k = 4$ *successful outcomes, can we reject the theory that predicts a successful rate of at most 10% at a significant level of* $\alpha = 1\,\%$ *?*

Classical approach

Binomial likelihood:

$$p(k\,|\,b, N) = Cb^k(1-b)^{N-k}$$

Classical inference with maximum likelihood for point estimate: $\mathrm{d}\ln p/\mathrm{d}b\,|_{b_0} = 0 \rightarrow b_0 = k/N,$

Recall the Taylor expansion about the maximum likelihood value:

$$\ln L(\theta) = \ln L(\theta_0) + \left.\frac{\mathrm{d}\ln L}{\mathrm{d}\theta}\right|_{\theta_0} \mathrm{d}\theta + \left.\frac{1}{2!}\frac{\mathrm{d}^2\ln L}{\mathrm{d}\theta^2}\right|_{\theta_0} \mathrm{d}\theta^2 + O(\mathrm{d}\theta^{3,4,\ldots})$$

Higher-order vanish in the case of Gaussian likelihood.

When N is large, the likelihood converges to Gaussian (under ***CLT***).

For Gaussian likelihood the quadratic term offers an exact solution to the uncertainty σ_b.

$$\sigma_b = \left(-\left.\frac{\mathrm{d}^2\ln p(b)}{\mathrm{d}b^2}\right|_{b_0}\right)^{-1/2} = \left[\frac{b_0(1-b_0)}{N}\right]^{1/2}$$

When N *is small, use proper Bayesian analysis!*

Bayesian Hypothesis Testing

We ask for a given observation of $N = 10$ *measurements with* $k = 4$ *successful outcomes, can we reject the theory that predicts a successful rate of at most 10% at a significant level of* $\alpha = 1\,\%$ *?*

Under *Gaussian approximation*,
$p(b < 0.1 \,|\, k = 4, N = 10) \sim 0.03$
therefore, the theory cannot be rejected at 1% significance level.

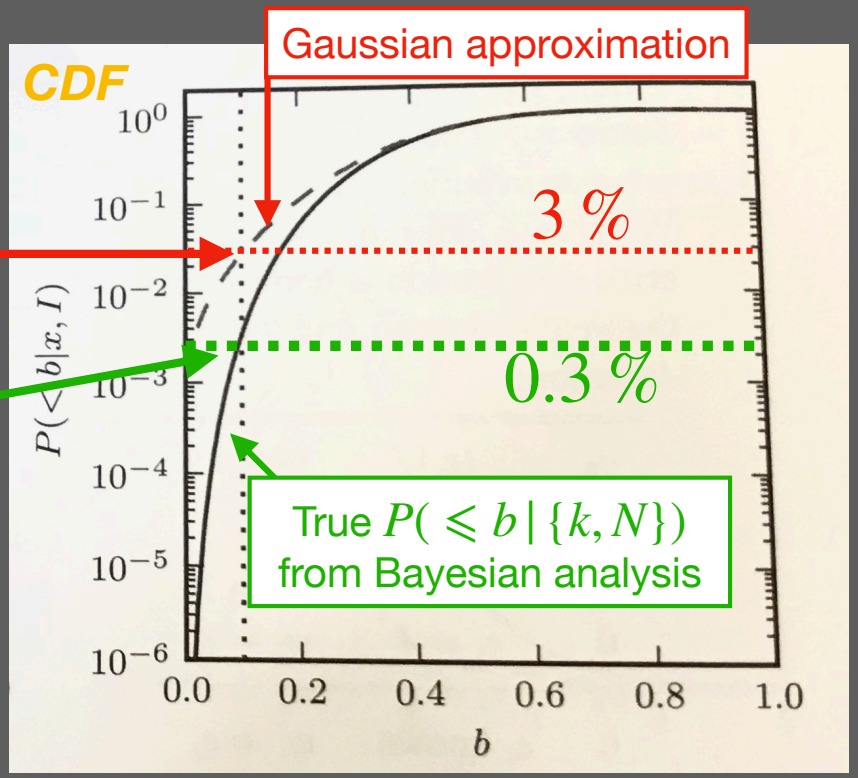
However, using *Bayesian* analysis, the theory will be ruled out at 1% significance level!

Homework is waiting for you ;)

Bayesian approach

the ***posterior*** probability for b:

$$p(b \,|\, k, N) = Cb^k(1-b)^{N-k}$$



Bayesian Model Selection

Question: "Which models are better supported by the available data?"

e.g., Gaussian vs. Cauchy distribution? Straight line vs. parabola?

Note: $\int p'(\vec{\theta}\,|\,D,M,I)\,\mathrm{d}\vec{\theta} = 1$ is normalization for $p'(\vec{\theta}\,|\,D,M,I)$ while doing parameter estimation given model M. While $\int p''(\vec{\theta},M\,|\,D,I)\,\mathrm{d}\vec{\theta} = p(M\,|\,D,I)$ with $p''(\vec{\theta},M\,|\,D,I) \neq p'(\vec{\theta}\,|\,D,M,I)$.

Model posterior

$$p(M\,|\,D,I) = \int p(\vec{\theta},M\,|\,D,I)\,\mathrm{d}\vec{\theta} = \frac{\int p(D\,|\,\vec{\theta},M,I)\,p(\vec{\theta},M\,|\,I)\,\mathrm{d}\vec{\theta}}{p(D\,|\,I)}$$

$$= \frac{\int p(D\,|\,\vec{\theta},M,I)\,p(\vec{\theta}\,|\,M,I)\,p(M\,|\,I)\,\mathrm{d}\vec{\theta}}{p(D\,|\,I)} = \frac{p(M\,|\,I)\int p(D\,|\,\vec{\theta},M,I)\,p(\vec{\theta}\,|\,M,I)\,\mathrm{d}\vec{\theta}}{p(D\,|\,I)}$$

$$= \frac{p(D\,|\,M,I)\,p(M\,|\,I)}{p(D\,|\,I)}$$

$E(M) \equiv p(D\,|\,M,I)$ is called the model's *evidence!*
It is the likelihood at the model
(instead of parameter) estimate level!

$$p(M\,|\,D,I) = \frac{E(M)\,p(M\,|\,I)}{p(D\,|\,I)}$$

Bayesian Model Selection

$$p(M \mid D, I) = \frac{E(M)\, p(M \mid I)}{p(D \mid I)}$$

The odds ratio:
(model posterior ratio)

$$Q_{21} \equiv \frac{p(M_2 \mid D, I)}{p(M_1 \mid D, I)} = \frac{E(M_2)}{E(M_1)} \frac{p(M_2 \mid I)}{p(M_1 \mid I)} = B_{21} \frac{p(M_2 \mid I)}{p(M_1 \mid I)} .$$

Bayes factor (model likelihood ratio, evidence ratio)

$Q_{21} > 100$ "decisive" evidence in favor of M_2
$Q_{21} > 10$ "strong" evidence in favor of M_2
$Q_{21} < 3$ "not worth more than a bare mention"

---

e.g., whether the data better follow a Gaussian distribution (M_1) or a Lorentzian distribution (M_2)? Q_{12} will be able to help with comparison between models.

Bayesian Model Selection

A special case of model comparison under Bayesian inference is to test (and reject) a ***null** hypothesis* M_1 ***against*** an ***alternative*** hypothesis M_2

Consider a simple case — coin tossing N ***times,*** k ***heads***

Instead of doing parameter estimation of success rate b*, I can compare between two hypotheses below to determine* ***whether*** *my coin flips head with a success rate of* b_**?*

- Hypothesis M_1: the coin has a ***known*** heads probability b_* i.e., with a prior $\delta(b - b_*)$
- Hypothesis M_2: the heads probability b follow a uniform prior of $b \in [0, 1]$

If M_2 *is a better model, then* M_1 *can be rejected.*

Binomial distribution

$$p(k \mid b, N) = \frac{N!}{k!(N-k)!} b^k (1-b)^{N-k}$$

$\bar{k} = bN$ (mean)

$\sigma_k = [Nb(1-b)]^{1/2}$ (standard deviation)

$$E(M) = \int p(D \mid \vec{\theta}, M, I) p(\vec{\theta} \mid M, I) \, \mathrm{d}\vec{\theta}$$

$$Q_{21} = \frac{E(M_2)\ p(M_2 \mid I)}{E(M_1)\ p(M_1 \mid I)}$$

$$Q_{21} = \int_0^1 \left(\frac{b}{b_*}\right)^k \left(\frac{1-b}{1-b_*}\right)^{N-k} \mathrm{d}\, b$$

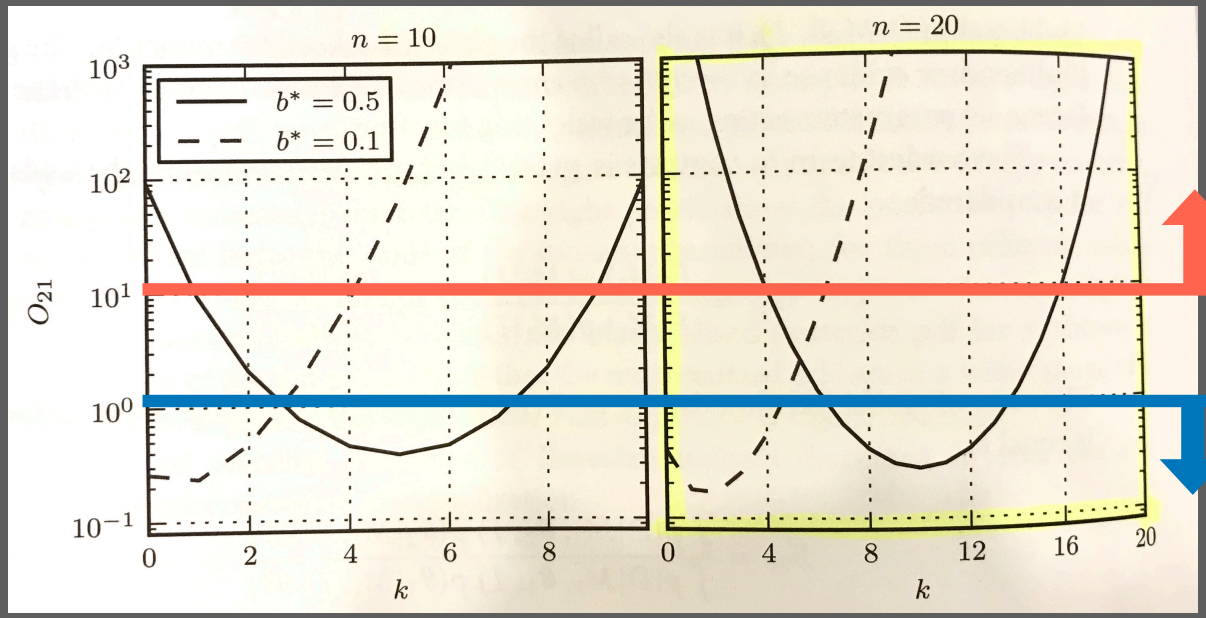
It is a function of data k, N and b_*.

Consider a simple case — coin tossing N times, k heads

- Hypothesis M_1: the coin has a ***known*** heads probability b_* i.e., with a prior $\delta(b - b_*)$
- Hypothesis M_2: the heads probability b is unknown - a uniform prior of $b \in [0, 1]$

$$Q_{21} = \int_0^1 \left(\frac{b}{b_*}\right)^k \left(\frac{1-b}{1-b_*}\right)^{N-k} \mathrm{d}\, b \qquad \approx \sqrt{2\pi}\sigma_b \left(\frac{b_0}{b_*}\right)^k \left(\frac{1-b_0}{1-b_*}\right)^{N-k}$$

Where $b_0 = k/N$, and $\sigma_b = \sqrt{b_0(1-b_0)/N}$.



M_2 winning over M_1,
M_1 is rejected. ✗

M_2 is not dominantly better!
M_1 cannot be rejected ✓

★ When k is very close to Nb_*, Q_{21} is small, M_1 cannot be rejected by data!

★ When the actual outcome k is very different from M_1's prediction Nb_*,
Q_{21} becomes large — M_1 is no longer favor by the data! *Reject the null hypothesis!*

#

Bayesian Model Selection

$$Q_{21} = \frac{E(M_2)\ p(M_2 \mid I)}{E(M_1)\ p(M_1 \mid I)} \qquad E(M) = \int p(D \mid \vec{\theta}, M, I)\, p(\vec{\theta} \mid M, I)\, \mathrm{d}\vec{\theta}$$

However, Q_{12} could be hard/expansive to compute!!
It can be simplified under certain assumptions about the likelihood/posterior.

Assuming a flat prior of θ in range Δ_θ and a *Gaussian likelihood* $\sim \mathcal{N}(\theta_0, \sigma_\theta)$,

$$E(M) \approx \sqrt{2\pi}\, L^0(M) \frac{\sigma_\theta}{\Delta_\theta}, \quad L^0(M) \equiv p_{\max}(D \mid \theta, M, I) = p(D \mid \theta_0, M, I)$$

With N data sets and k parameters, the evidence is given by:

$$E(M) \approx \sqrt{2\pi}\, L^0(M) \Pi_{j=1}^{k} \frac{\sigma_{\theta_j}}{\Delta_{\theta_j}},$$

When the data are more informative than the prior, I.e., $\sigma_\theta \ll \Delta_\theta$, every inclusion of a (new) model parameter θ makes $E(M')$ penalized by $\sigma_\theta / \Delta_\theta$. In this sense, the model with less parameters tends to win (*"Occam's Razor"*), *unless* new model significantly increases data likelihood $L^0(M')$…

Bayesian Model Selection

If NOT, the best statistic to use is the odds ratio!

Under the assumption of *Gaussian* likelihood/posterior pdf, with N data sets and k parameters, the evidence is given by:

$$E(M) \approx \sqrt{2\pi}\, L^0(M) \Pi_{j=1}^{k} \frac{\sigma_{\theta_j}}{\Delta_{\theta_j}}, \qquad \textit{The larger the better!}$$

If so, BIC can be used!

Bayesian Information Criteria *(BIC, Schwarz criterion)**:***

$$BIC \propto -\ln[E(M)] \qquad \textit{The smaller the better!}$$

Consistent with AIC!

$$BIC \equiv -2\ln[L^0(M)] + k \ln N$$

Prefer bigger *Prefer smaller*

k: number of parameters

N: number of data points

$\sigma_\theta \propto 1/\sqrt{N}$ [CLT]

★ *BIC* is easier to compute compared to the *odds ratio* Q_{21}

Homework is waiting for you ;)

What is Bayesian statistical inference?

Bayes' theorem

The Prior (joint probability) ***Parameter*** ***Model***

$$p(M, \vec{\theta} \mid I) = p(\vec{\theta} \mid M, I)\ p(M \mid I)$$

Likelihood function

Of an initial model $M[\theta]$ that generate observed data D, under the *prior information* I

Prior - "a priori joint probability" given *prior information* I, and in the absence of any *new* data!

Posterior pdf of an improved model $M[\theta]$ given observed data and prior information

$$p(M, \vec{\theta} \mid D, I) = \frac{p(D \mid M, \vec{\theta}, I)\ p(M, \vec{\theta} \mid I)}{p(D \mid I)}$$

Model M described by parameter θ

Given data and *prior information*

Symmetrize the probability calculation between data and model/parameter!

Move from old knowledge/constraints on model to a new knowledge/constraints

Review

Bayesian

To find best model parameters:

Maximize either the joint posterior $p(\vec{\theta}\,|\,D)$ or the marginalized posterior $p(\theta\,|\,D)$ to yield the ***maximum a posteriori (MAP)*** estimate

Or

Obtain the ***posterior mean*** (or median):

$$\bar{\theta}_1 = \int \theta_1\, p(\theta_1\,|\,D)\, \mathrm{d}\theta_1,$$

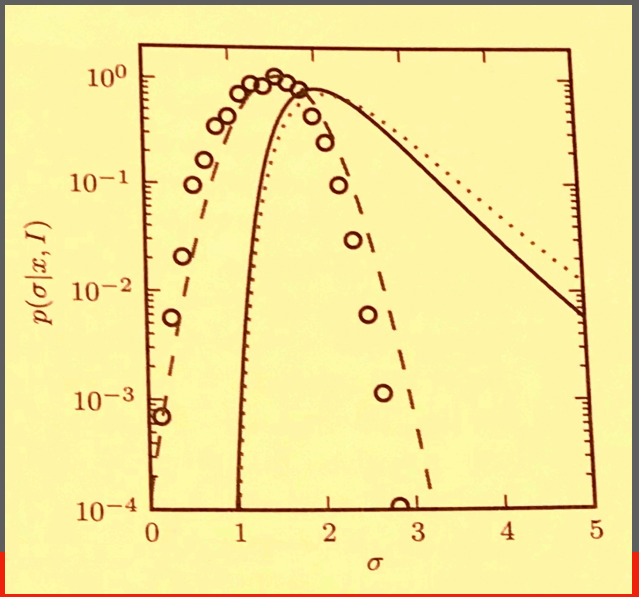
where $p(\theta_1\,|\,D)$ is obtained using marginalization - integration of $p(\vec{\theta}\,|\,D)$ over all *other* model parameters except θ_1.
Don't forget to renormalize:

$$\int p(\theta_1\,|\,D)\, \mathrm{d}\theta_1 = 1$$

Review

To quantify parameter uncertainties
— Credible regions:

Analytically calculating the posterior interval or sampling the posterior distribution numerically (e.g., MCMC).



When N is small (i.e., $N < 100$ for most applications), go Bayesian!
The classical estimates under the ***CLT (e.g., assuming Gaussian statistics or resampling-based methods)*** fail to predict the correct parameter pdf!

Bayesian

Parameter estimate given model:

$$p(\vec{\theta}\,|\,D,M,I) = \frac{p(D\,|\,\vec{\theta},M,I)\,p(\vec{\theta}\,|\,M,I)}{E(M)}$$

Model estimate/comparison:

$$p(M\,|\,D,I) = \frac{E(M)\,p(M\,|\,I)}{p(D\,|\,I)}$$

$$E(M) \equiv p(D\,|\,M,I) = \int p(D\,|\,\vec{\theta},M,I)p(\vec{\theta}\,|\,M,I)\,\mathrm{d}\vec{\theta}$$

$E(M) \equiv p(D\,|\,M,I)$ is called the model's *evidence!*

- ★ Denominator (normalization-keeping) at parameter estimate level;
- ★ Also the likelihood at the model estimate level!

Model comparison through the odds ratio Q_{12} ~ given through the Bayes factor B_{21} — the ratio between two model evidence $E(M)$s.

Or, when *Gaussian* likelihood is a good approximation, *BIC* can be used.

Bayesian Information Criterion $BIC \equiv -2\ln[L^0(M)] + k\ln N$

Aikake Information Criterion $\mathrm{AIC} \equiv -2\ln[L_0(M)] + 2k + \frac{2k(k+1)}{N-k-1}$

- ★ *BIC* is easier to compute compared to the *odd ratio* Q_{21}
- ★ *BIC* can penalize complex models more than the *AIC.*

Review

Bayesian versus Frequentist

- ★ Possible to live completely in either paradigm as a data analyst!
- ★ Symmetry and grand unification between parameter and data in Bayesian!
- ★ Bayesian allows extra information: with prior information, make useful conclusion using small sample, and estimate measurement errors!
- ★ Bayesian computation expansive but (numerically) straightforward!

We are not done with Bayesian!!

- ★ *Monte Carlo* methods for Bayesian Inference
- ★ Further applications in *model regression*