Introduction to (mostly Bayesian) statistics

Martin Kunz
Université de Genève

Overview

- Probability distributions
- Bayes theorem
- Parameter estimation and model selection
- Practical aspects
  - Gaussians
  - Fisher matrix / error forecasts
  - MCMC

Probability distribution(s)

- Space of Results Ω (e.g. coin: Ω = {🡅,🡇})
- Probability measure P: $P(A) \geq 0$, $P(\Omega)=1$, $P(A_1+A_2)=P(A_1)+P(A_2)$ for A1, A2 disjoint
- Random variable X : Ω -> R (e.g. coin: X(🡅)=1)
- Probability density function (pdf): $P(x) = \text{prob}(X=x)$ -> $P(x) \geq 0$, $\Sigma_x P(x) = 1$
- Cumulative distribution function (cdf): $F(x)=\text{prob}(X \leq x)$ -> $F(x)=\Sigma_{u \leq x} P(u)$
- Joint distribution: $P(x,y)=\text{prob}(X=x \text{ AND } Y=y)$
- Marginal distribution: $P(x) = \text{prob}(X=x) = \Sigma_y P(x,y)$ (and the same for y)
- Conditional distribution: $P(x|y) = \text{prob}(X=x \text{ IF } Y=y)$
- Theorem: $P(x,y) = P(x|y) P(y) = P(y|x) P(x)$
- Expectation value: $E[g(X)] = \Sigma_x g(x) P(x)$

mean, variance, etc

- Mean: $\mu=E[X]=\Sigma_x \; x \; P(x)$ -> $E[cX] = c \; E[x]$
- Variance $\sigma^2=E[X^2]-E[X]^2=\Sigma_x \; (x-\mu)^2 \; P(x)$
  -> $\sigma^2[cX] = c^2 \; \sigma^2[X]$
- Covariance $Cov(X,Y) = \Sigma_{x,y} \, (x-\mu_x)(y-\mu_y) \; P(x,y)$
- $Cov(X,Y) = E[XY] - \mu_x\mu_y$

- X,Y independent <-> $P(x,y) = P(x) \; P(y)$
  -> $P(x|y) = P(x,y)/P(y) = P(x)$
  and $Cov(X,Y) = 0$

- $\sigma^2[X\pm Y] = \sigma^2[X] + \sigma^2[Y] \pm Cov(X,Y)$

Normal (Gaussian) pdf

- Normal distribution:

$$\boxed{P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \mathcal{N}(\mu, \sigma^2)}$$

- mean: μ, variance: σ^2
- $Z = (X-\mu)/\sigma$ reduced variable, $P(z) = N(0,1)$
- Generic limiting case (central limit theorem)
- If $X_1, X_2, \ldots, X_n$ indep. $N(0,1)$: $\chi^2 = \Sigma_i X_i^2$ has the so-called chi-squared distribution with n degrees of freedom
- For χ^2: mean n, variance 2n

more on Normal pdf

- Gaussian pdf is also 'least informative' (maximum entropy) choice if only mean and variance known
- In reality, often exponential decrease at high x/σ is too steep, 'heavy tails'
- Generalisation for vector of random variables $X=(X_1,X_2,\ldots,X_n)$: multivariate Gaussian

$$\boxed{P(x) = \frac{1}{(2\pi|C|)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i,j=1}^{n}(x_i-\mu_i)C_{ij}^{-1}(x_j-\mu_j)\right]}$$

  - given by mean vector μ and covariance matrix C (symmetric, positive -> eigenvalues are real & positive)
  - if X_i independent: $C=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$ and

$$P(x) = \prod_{i=1}^{n} P(x_i)$$ product of univariate pdf's

Statistics

- Typical case: Data $D=\{(x_i,y_i,\sigma_i)\}$ [σ: error on y]
- Assumption: $P(y_i|x_i,y(x),\sigma_i)= N(y(x_i),\sigma_i^2)$ indep.
- In general y(x) is a function of parameters θ, e.g. $y(x) = a*x+b$ -> $\theta=\{a,b\}$

$\Rightarrow$ define $\chi^2 = \sum_i \frac{[y_i - y(x_i;\theta)]^2}{\sigma_i^2} \quad \rightarrow P(D|\theta) \propto e^{-\chi^2/2}$

χ^2 has chi-square distribution with ν= (# data points) - (# parameters) degrees of freedom

- best fit at $\frac{\partial \chi^2}{\partial \theta_j} = 0$ ('maximum likelihood', ML)
- can check 'goodness of fit' of minimal χ^2
- Taylor expansion of at χ^2 ML -> $H_{jk} \equiv \frac{1}{2}\frac{\partial^2 \chi^2}{\partial \theta_j \theta_k}$ -> $Cov(\theta_j,\theta_k)=(H^{-1})_{jk}$
- $P(D|\theta)$ only normal in θ if model $y(x;\theta)$ is linear in θ!

Bayesian statistics

- In general we want to know the underlying parameters θ, i.e. P(θ|D), not P(D|θ)
- P(θ|D) has no probabilistic interpretation in a frequentist sense: the parameters θ are not random variables
- Bayesian interpretation: 'limited knowledge'
- Formally just application of Bayes theorem:

$$P(D,\theta) = P(D|\theta)P(\theta) = P(\theta|D)P(D) \Rightarrow \boxed{P(\theta|D) = P(D|\theta)\frac{P(\theta)}{P(D)}}$$

- Mathematical proofs exist that construction is at least self-consistent (cf eg Cox theorem)

Bayes theorem example

- You have a mind-scanner that can identify a terrorist with 99.99% probability and gets it wrong in only 0.01% of cases
- 1 in 1’000’000 is a terrorist
- should you shoot people who fail the mind-scanner test?

Bayes theorem example

- You have a mind-scanner that can identify a terrorist with 99.99% probability and gets it wrong in only 0.01% of cases
- 1 in 1’000’000 is a terrorist
- should you shoot people who fail the mind-scanner test?

X: is a terrorist, Y: fails mind-scanner

$P(Y|X) = 0.9999$

$P(X|Y) = P(Y|X)P(X)/P(Y) \sim 1*10^{-6}*10^{4} \sim 10^{-2}$!!!

Parameter estimation

- $P(D|\theta)$: likelihood $L(\theta)$ -> 'given' by experiment
- $P(\theta|D)$: posterior -> that's what we want
- $P(\theta)$: prior [$P(D)$: left for later]
- Prior: necessary, measure on parameter space, typical choices:
  - $P(\theta)$ constant -> 'flat prior', $P(D|\theta) \sim L(\theta)$
  - $P(\theta) \sim 1/\theta$ -> prior flat in $\log(\theta)$ -> no scale for θ

  (there is a whole literature on how to choose priors)
- What to estimate?
  - Mean & error: $\mu_\theta = \Sigma_\theta\, \theta\, P(\theta|D)$, $C(\theta_i,\theta_j)$ [as before]
  - Maximum: $\max_\theta P(\theta|D)$ -> max. likelihood for flat prior
  - 'credible regions', e.g. 95% parameter volume

Explicit example

Very simple example:

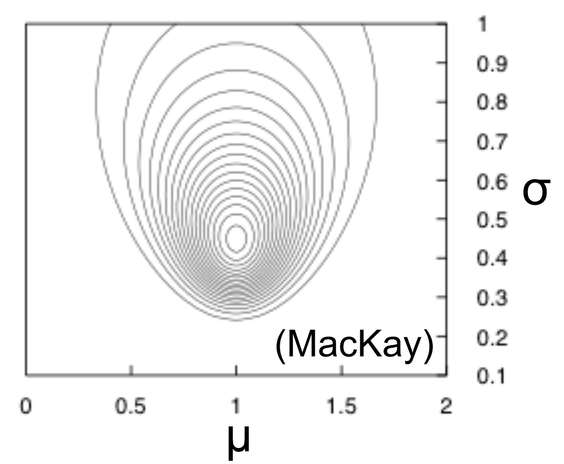
- D = {x_i, i=1,…,n} drawn indep. from N(μ,σ^2)
- Estimate μ and ln σ

1. Priors: P(μ)=const, P(ln σ) = const
2. Likelihood: product of P($x_i|\mu,\sigma^2$) over all points

$$P(D|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{n(\mu-\bar{x})^2 + nS^2}{2\sigma^2}\right\}$$

$\left(n\bar{x} = \sum_i x_i,\ nS^2 = \sum_i (x_i-\bar{x})^2\right)$ sufficient statistics



3. Posterior: P(μ, ln $\sigma|D$) ~ P($D|\mu$,ln σ)

Explicit example II

1. Maximum of posterior = maximum of likelihood, it is at $\{\mu = \bar{x}, \sigma = S\}$ (compute dL/dθ=0)
2. Assume σ known -> want P(μ|D,σ)

$$P(\mu|\{x_i\}_{i=1}^n, \sigma) \propto \exp\left\{-\frac{n(\mu-\bar{x})^2}{2\sigma^2}\right\}$$

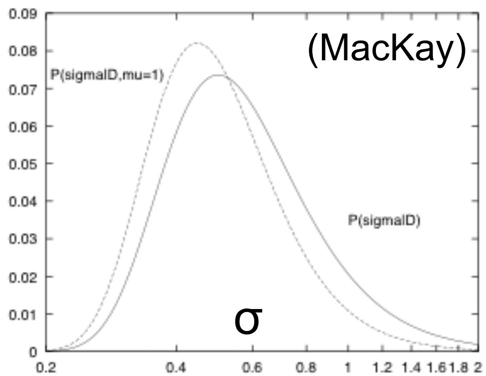
$$\rightarrow P(\mu) = \mathcal{N}(\bar{x}, \sigma^2/n)$$

3. Assume both μ and σ unknown, what is P(σ|D)?

$$P(D|\sigma) = \int P(D,\mu|\sigma)d\mu = \int P(D|\sigma,\mu)P(\mu)d\mu$$

Gaussian integral for P(μ)=const, can be done, now maximum at

$$\sigma^2 = \frac{n}{n-1}S^2$$



Explicit example III

4. Both μ and σ unknown (as 3), what is P(μ|D)?

$$P(\mu|D) = \int P(\mu, \sigma|D) d\sigma \propto \int_0^\infty \sigma^{-(n+1)} \exp\left\{ -\frac{n(\mu-\bar{x})^2 + nS^2}{2\sigma 2} \right\} d\sigma$$

can be solved e.g. by setting $u = A/\sigma^2$

$$\rightarrow P(\mu|D) \propto A^{-n/2} \propto 1/\left(n(\mu-\bar{x})^2 + nS^2\right)^{n/2}$$

(normalisation e.g. from $\int d\mu P(\mu|D) = 1$)

-> Student's t distribution [notice heavy tails!]

(here resulting from superposing Normal distributions with different widths)

-> this is the pdf to use when variance unknown!

Model selection

- So far we always assumed model to be known.
- If not, then we can add overall dependence on M

$$P(\theta|D,M) = P(D|\theta,M)\frac{P(\theta|M)}{\boxed{P(D|M)}}$$

- we want to know P(M|D)
- Bayes again: P(M|D) = $\boxed{\text{P(D|M)}}$ P(M) / P(D)
- And $$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(M_1)}{P(M_2)}\frac{P(D|M_1)}{P(D|M_2)} = \frac{P(M_1)}{P(M_2)}\boxed{B_{12}}$$

Bayes factor (absolute value of P(D|M) not so instructive)

- Since $\int P(\theta|D,M)d\theta = 1$

$$\boxed{P(D|M) = \int d\theta P(D|\theta,M)P(\theta|M)}$$

(likelihood used as f(θ) but normalised wrt D!)

goodness of fit vs model selection

250 coin tosses: 140 heads, 110 tails (<- D)

Random or not?

Likelihood: binomial $P(n_h, n_t|p) = \frac{(n_h + n_t)!}{n_h! n_t!} p^{n_h} (1-p)^{n_t}$

coin unbiased: p=1/2 => $P(n_h \geq 140|p=1/2) \sim 0.033$

-> looks bad!

goodness of fit vs model selection

250 coin tosses: 140 heads, 110 tails (<- D)
Random or not?

Likelihood: binomial $P(n_h, n_t|p) = \frac{(n_h + n_t)!}{n_h! n_t!} p^{n_h} (1-p)^{n_t}$

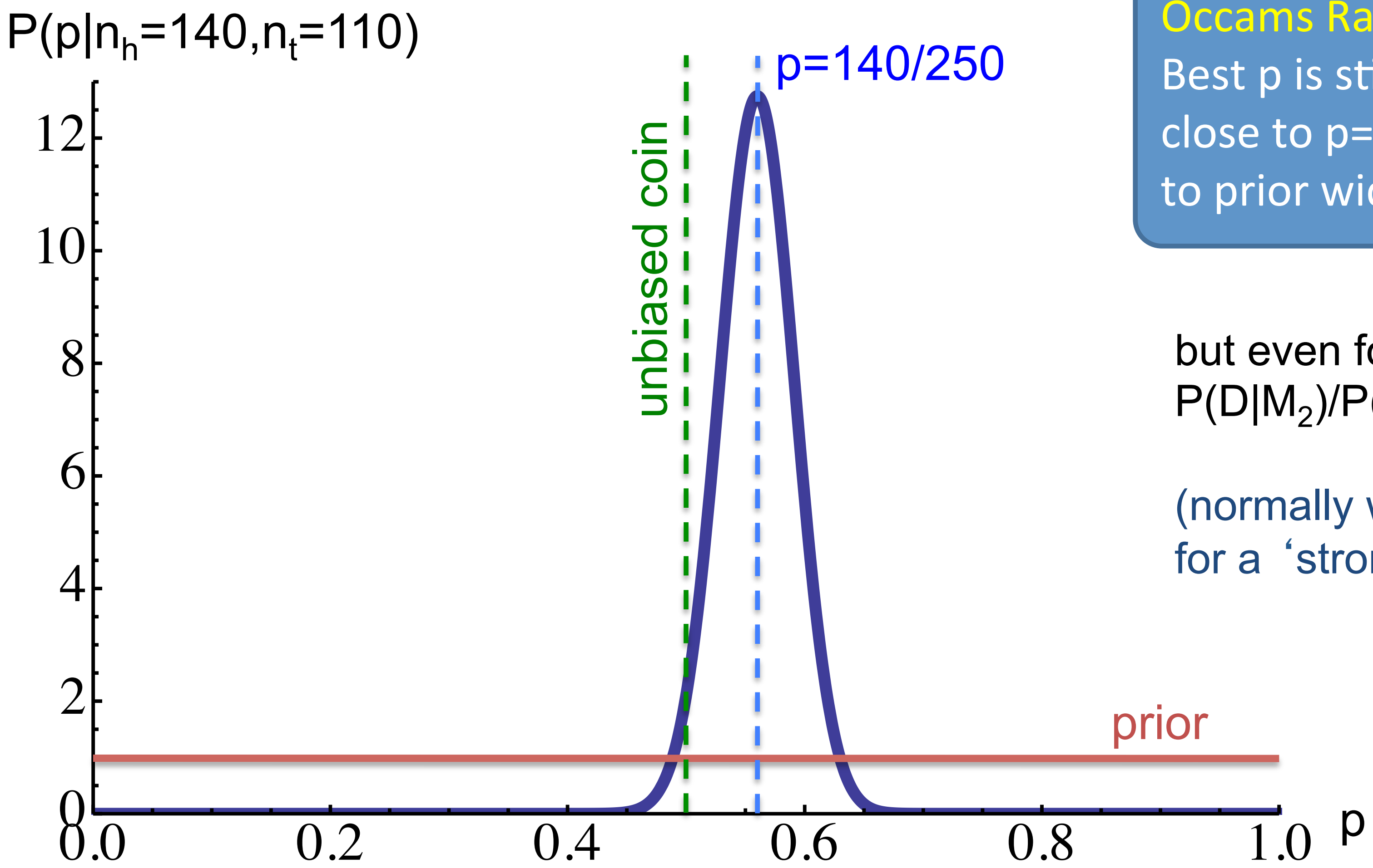
coin unbiased: p=1/2 => $P(n_h \geq 140 | p=1/2) \sim 0.033$
-> looks bad!

Bayes: M_0: p=1/2, M_1: p free parameter, P(p) uniform in [0,1]

$$\left.\begin{array}{l} P(D|M_0) \propto 1/2^{n_h+n_t} \\ P(D|M_1) \propto \int_0^1 dp p^{n_h} (1-p)^{n_t} = \frac{n_h! n_t!}{(n_h+n_t+1)!} \end{array}\right\} \quad \frac{P(D|M_1)}{P(D|M_0)} \approx 0.48$$

-> bad absolute goodness of fit should make you suspicious, but still need to find a better model!

model selection



Occams Razor:
Best p is still surprisingly close to p=1/2, relative to prior width.

but even for M_2: p=0.56
$P(D|M_2)/P(D|M_0) \sim 6.1$

(normally want B >> 10 for a 'strong' result)

Practical aspects

Often 10+ parameters (sometimes much more!)

Grid with 5 points on each side: $5^{10} \sim 10^7$ points

-> how to deal with high-dimensional spaces?

- Analytical approximation: Gaussians
- Numerical methods: MCMC

We would like a simple way to forecast accuracy of future experiments

- Fisher matrix formalism
- (or just create a fake likelihood and analyze it)

Gaussians

Often likelihood / posterior is also approximately Gaussian in parameters -> Taylor expansion:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \frac{1}{2}\sum_{ij}(\theta_i - \hat{\theta}_i)\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}(\theta_j - \hat{\theta}_j) + \dots$$

Here peak $\hat{\theta}$ and a bit loosely $C_{ij}^{-1} = -\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}$ at peak

This is just proportional to a Gaussian / Normal multivariate pdf for the parameters θ:

$$P(\theta|C,\mu) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp\left\{-\frac{1}{2}(\theta-\mu)^T C^{-1}(\theta-\mu)\right\}$$

(In general a Gaussian pdf for the data [-> χ^2] does not imply a Gaussian pdf for the parameters, only if the model y(x;θ) is linear! But: central limit theorem!)

Gaussians

Big advantage:

- Products of Gaussians are Gaussians

$$\mathcal{N}(x;\mu_1,C_1)\mathcal{N}(x;\mu_2,C_2) = A_3\mathcal{N}(x;\mu_3,C_3)$$

$$C_3 = (C_1^{-1} + C_2^{-1})^{-1}, \quad \mu_3 = C_3(C_1^{-1}\mu_1 + C_2^{-1}\mu_2)$$

$$A_3 = \mathcal{N}(\mu_1;\mu_2,C_1+C_2)$$

- We can evaluate Gaussian integrals
  - Simple explicit marginalisation:

    marginal distribution is again Gaussian

$$\int \mathcal{N}(x_1,\ldots,x_q,\ldots,x_n;\mu,C)dx_1...dx_q = \mathcal{N}(x_{q+1},\ldots,x_n;\overline{\mu},\overline{C})$$

$\overline{\mu} = (\mu_{q+1},\ldots,\mu_n)$ and $\overline{C}$ is just the [q+1,n] submatrix of C

  - Compute model probabilities, etc
  - Fisher matrix formalism

Errors for Gaussians

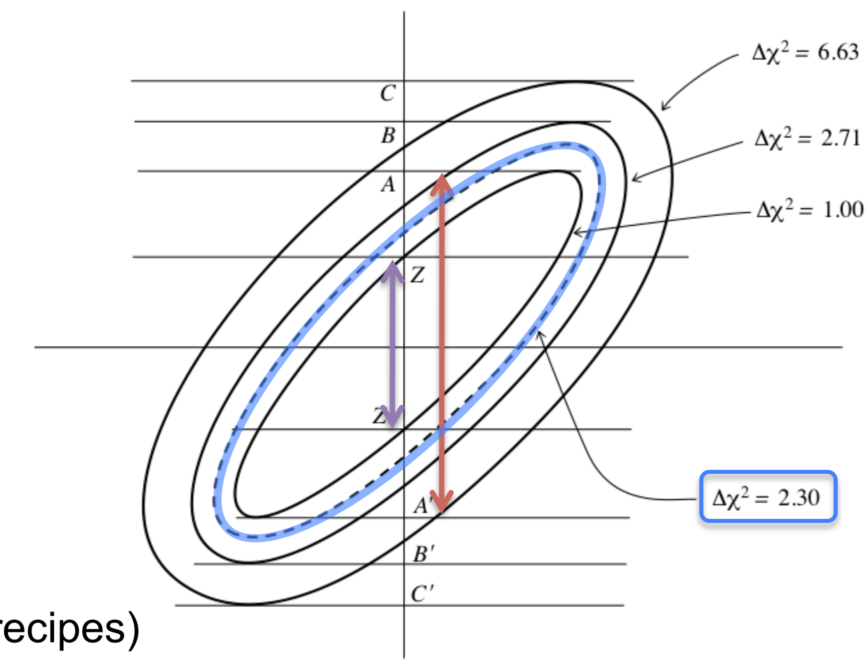
- Errors given by covariance matrix C = H^{-1}

$$H_{ij} \simeq -\frac{\partial^2 \ln P(\theta|D)}{\partial\theta_i \partial\theta_j} \qquad \Delta\chi^2 = \sum_{ij} (\theta_i - \hat{\theta}_i) H_{ij} (\theta_j - \hat{\theta}_j)$$

- Inverse of sub-matrix of H: conditional errors
- sub-matrix of inverse of H: marginal errors
- Constant χ^2 boundaries:
  Gaussian approximation!

$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom

	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8



(numerical recipes)

Fisher matrix formalism

- Fisher information matrix: measures information about parameters θ_i, defined as var(score), or

$$F_{ij} = \langle H_{ij} \rangle = -\left\langle \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right\rangle$$

- Expectation is taken over data realizations for given (fixed) model and 'fiducial' parameters
- Inverse of Fisher matrix can be seen as 'lower bound' on covariance matrix (Cramer-Rao bound)
- All results for Gaussians also apply here
- Due to expectation value, we don't need actual data realizations, only the specification of the experiment

Calculating Fisher matrices

- Explicit computation… simple form for normal data:

$$F_{ij} = \left(\partial_{\theta_i}\mu^T C^{-1}\partial_{\theta_j}\mu\right) + \frac{1}{2}\mathrm{tr}\left(C^{-1}\partial_{\theta_i} C C^{-1}\partial_{\theta_j} C\right)$$

- If you have a set of observables O_k and know the (expected) errors σ_k on O_k, then you can do error propagation:

$$F_{ij} = \sum_k \frac{\partial O_k}{\partial \theta_i}\frac{1}{\sigma_k^2}\frac{\partial O_k}{\partial \theta_j}$$

  - this generalizes in the obvious way to a covariance matrix for the O_k
  - If you have relative errors $\delta_k = \sigma_k/O_k$ then

$$F_{ij} = \sum_k \frac{\partial \ln O_k}{\partial \theta_i}\frac{1}{\delta_k^2}\frac{\partial \ln O_k}{\partial \theta_j}$$

simple Fisher example

Let’s revisit the simple Gaussian example:

$$L(\mu,\sigma) = P(D|\mu,\sigma) \; = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{n(\mu-\bar{x})^2 + nS^2}{2\sigma^2} \right\}$$

second derivatives of ln(L) and expectation:

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} \to -\frac{n}{\sigma^2} \qquad \frac{\partial^2 \ln L}{\partial \sigma \partial \mu} = \frac{n}{\sigma^3}\,(2(\mu-\bar{x})) \to 0$$

$$\frac{\partial^2 \ln L}{\partial \sigma^2} = \frac{n}{\sigma^4}\left(\sigma^2 - 3S^2 - 3(\mu-\bar{x})\right) \to -2\frac{n}{\sigma^2}$$

- The Fisher matrix is diagonal → errors independent
- error on μ: σ/√n, error on σ: σ/√(2n)
- no actual data realization is required
- the true posterior of σ is non-Gaussian

Markov-Chain Monte Carlo

Aim: create ensemble of parameter samples $\{\theta^{(i)}\}$ that are drawn from posterior pdf, i.e.

$$P(\theta|D) \sim 1/N \, \Sigma_i \, \delta(\theta-\theta^{(i)})$$

-> expectation values: $\langle g(\theta)\rangle \sim 1/N \, \Sigma_i \, g(\theta^{(i)})$

-> marginalisation becomes projection, just drop the parameters that you want to marginalise

-> credible region: find volume enclosing x% of points (marginalise first for less dimensions)

Most popular algorithm: Metropolis-Hastings

Metropolis-Hastings

0. init: choose random point x in parameter space
1. step: choose new point y from proposal distribution $q(y|x)$
2. test: accept new point with probability $\min[1,P(y)/P(x)]$ (*)
3. if accepted set $x=y$
4. store x (even if not changed!), go to 1 and repeat

(*) this condition assumes symmetric proposal distribution, $q(y|x) = q(x|y)$ otherwise acceptance prob. slightly more complicated, $\min[1,\{P(y)q(y|x)\}/\{P(x)q(x|y)\}]$.

- ***Burn-in:*** initial period, should be discarded
- ***Convergence:*** need to collect samples until we have a fair sample of target distribution, this can be difficult to judge (impossible in general). Diverse criteria exist.

Metropolis-Hastings II

In theory the algorithm converges independently of the choice of proposal distribution $q(x|y)$, in reality this tends to be the most important choice.

Usual choice is 2.3*Gaussian centered on x with parameter covariance matrix (-> rotated ellipsoid).

Of course to do this one needs to know the answer -> re-compute covariance matrix on the fly, but in principle need to fix it for samples used in analysis.

small project

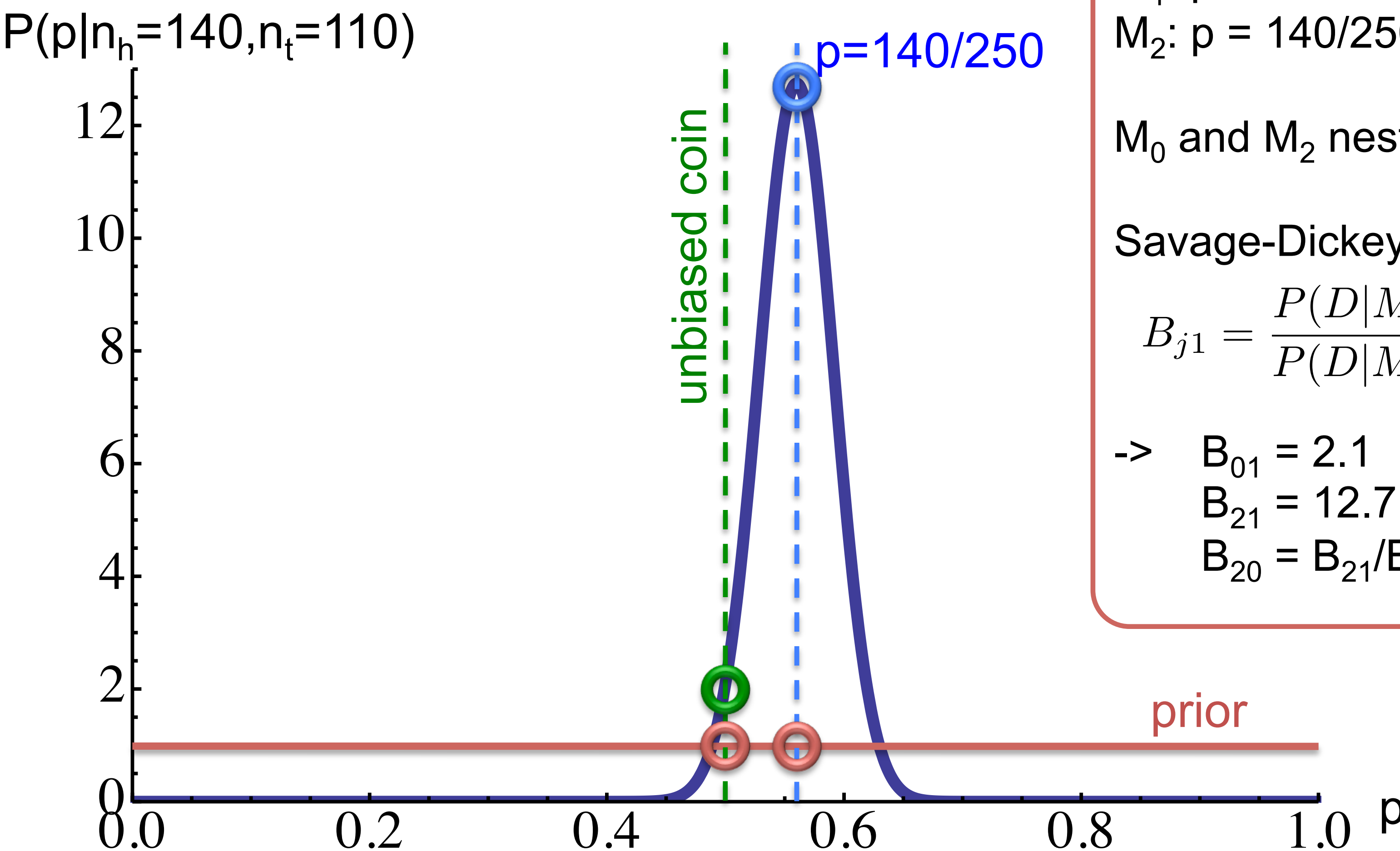
- get (simulated) data $[x_i, y_i, \sigma_i]$ from here:
  http://mpej.unige.ch/~kunz/poly_stat.dat.gz
- model: $y(x)=a_0+a_1x+a_2x^2$
- y_i are Gaussian around $y(x_i)$ with error σ_i
- write a little MCMC program to find parameters and correlations
- check by computing (semi-analytically) $d\chi^2/da_i = 0$ [easy for linear models]
- can also try model-comparison to check models $y(x) = \Sigma_i\, a_i\, x^i$ for different i_{max}

Practical model selection

The integration over (Likelihood)x(prior) is normally hard, MCMC chains are not good enough.

- Numerical methods: thermodynamic integration, nested sampling
- Use Gaussian approximation (possibly with several Gaussians: mixture models)
- For nested models (the simpler model is same as general model with some parameters fixed) *Savage-Dickey*: Bayes factor is just posterior/ prior of general model at nested point, marginalised over all common parameters.

Savage-Dickey example



M_0: $p = ½$
M_1: p free
M_2: $p = 140/250$

M_0 and M_2 nested in M_1

Savage-Dickey:

$$B_{j1} = \frac{P(D|M_j)}{P(D|M_1)} = \frac{P(p|D)}{P(p)}$$

-> $B_{01} = 2.1$
$B_{21} = 12.7$
$B_{20} = B_{21}/B_{01} = 6.1$

Summary

- Bayes: P(θ|D) ~ P(D|θ) P(θ)
- Prior is an integral part of method (but posterior not very sensitive to it if data is any good)
- Bayesian statistics allows for (relatively) straightforward manipulation of probabilities
- Non-trivial examples tend to need MCMC or Gaussian approximations
- Model selection: P(M|D)
- Bayes factor $B_{01}=P(D|M_0)/P(D|M_1)$ ('betting odds')
- want |ln(B)| > 2-3 for strong results
- Model selection is much more sensitive to prior