COSMOLOGICAL DATA ANALYSIS

中山大学

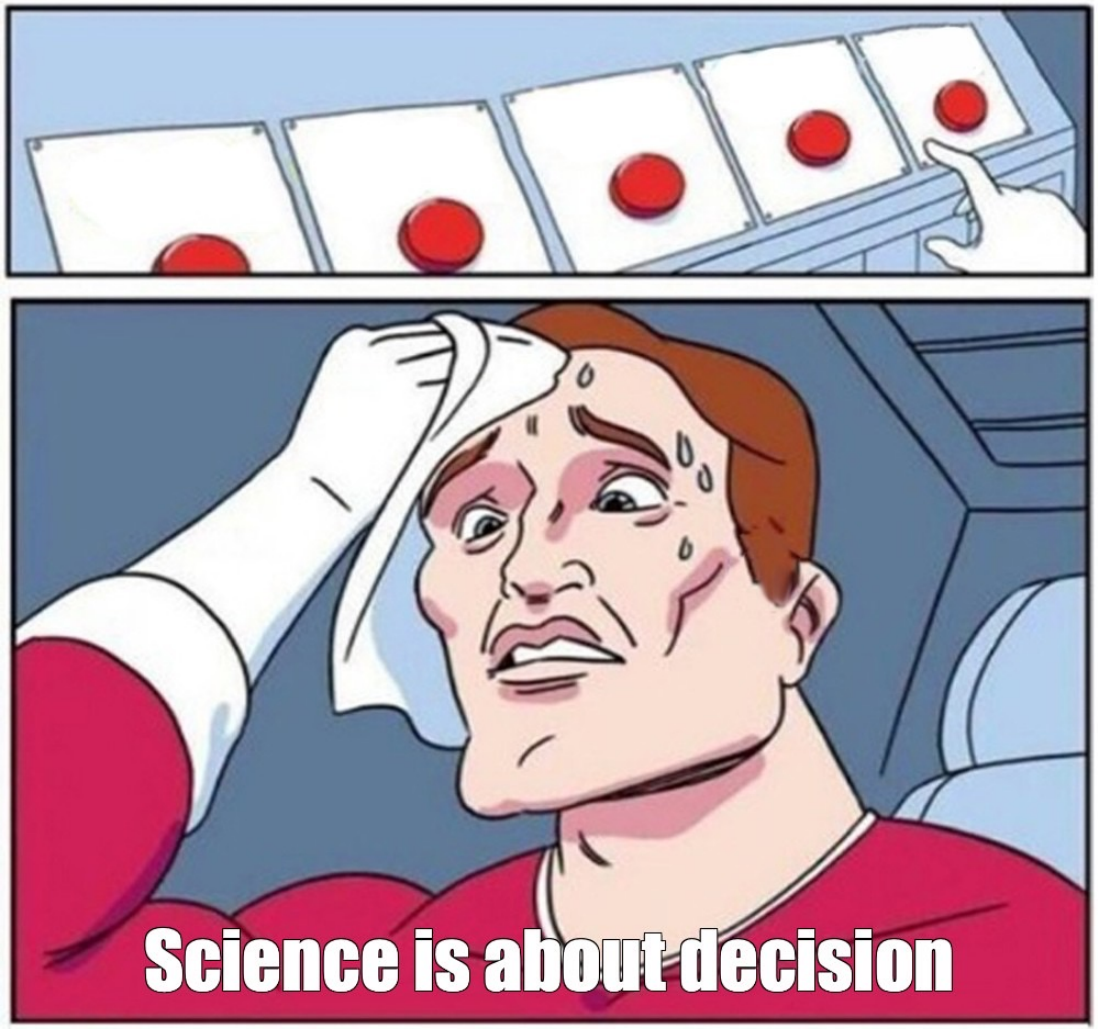
王 鑫

CONTENTS

BASED ON ...

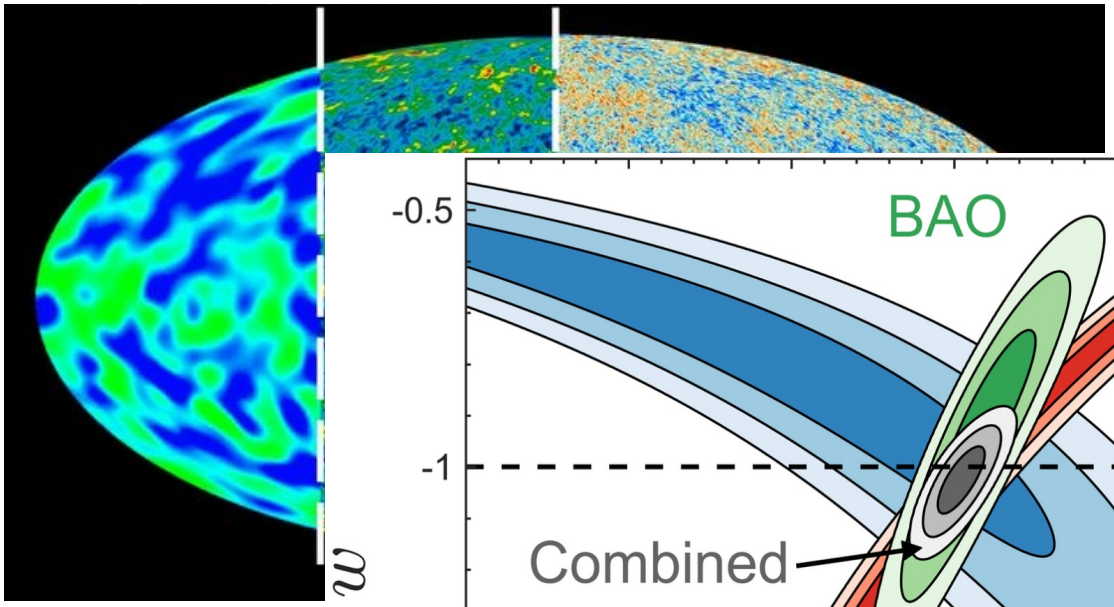
- ***Practical statistics for Astronomer**, Cambridge University Press, J. V. Wall and C. R. Jenkins*
- ***Modern Cosmology**, Second Edition, Academic Press, Scott Dodelson, Fabian Schmidt*
  - *Chapter 14. "Analysis and inference"*
- ***Information Theory, Inference, and Learning Algorithms**, Cambridge University Press, David J.C. MacKay*

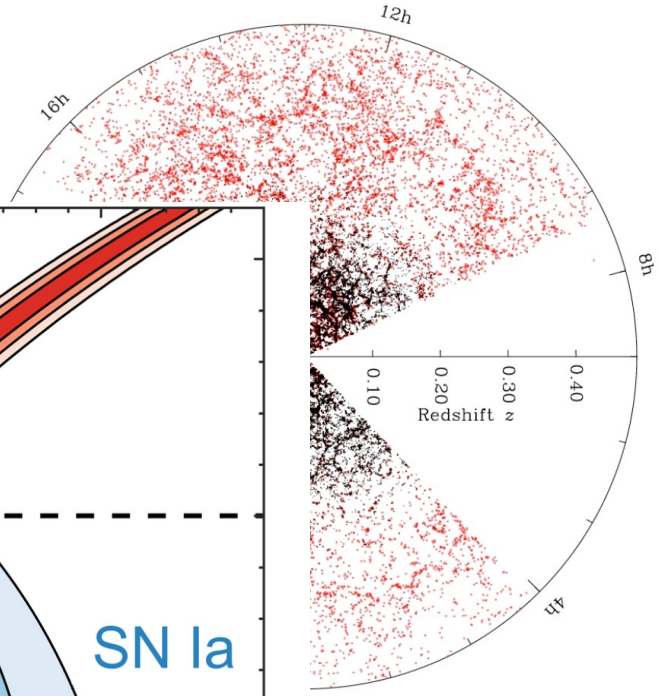
SCIENCE IS ABOUT DECISION

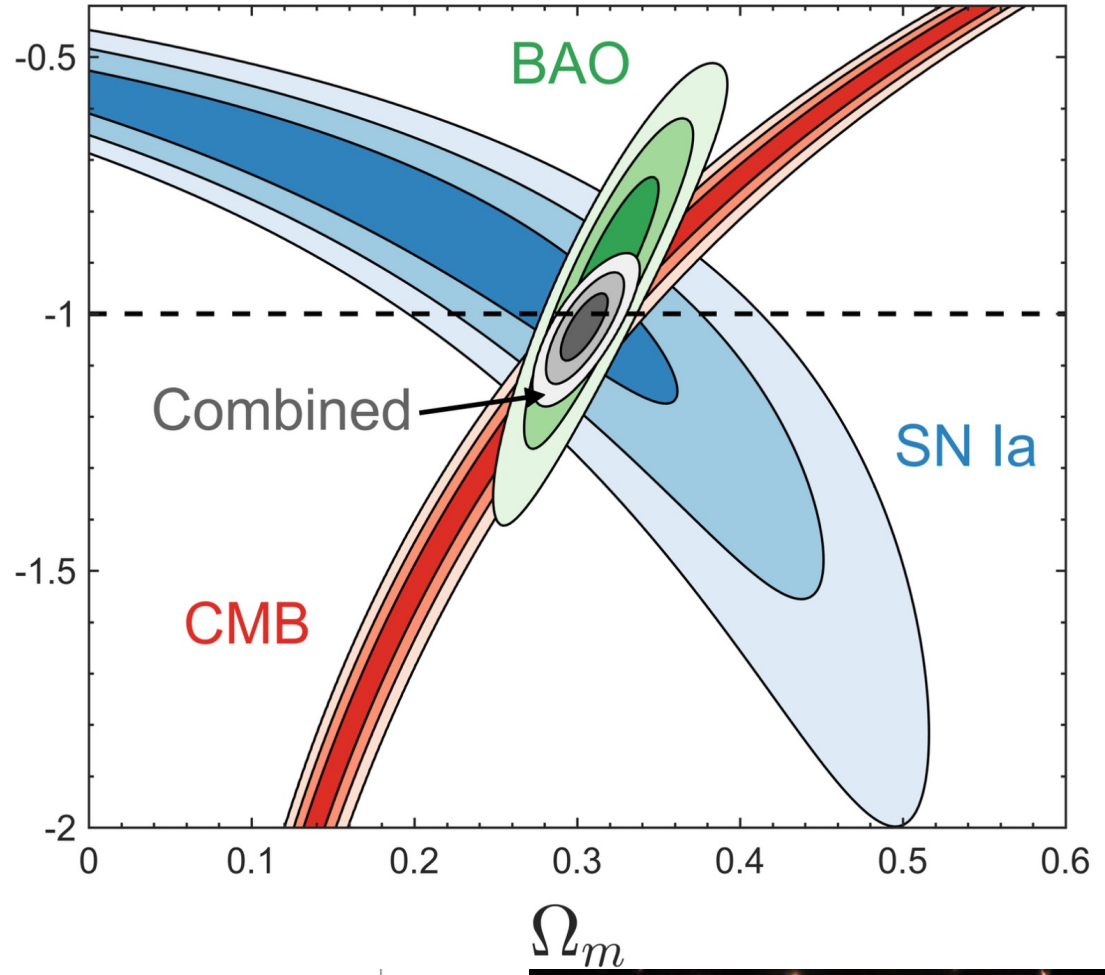


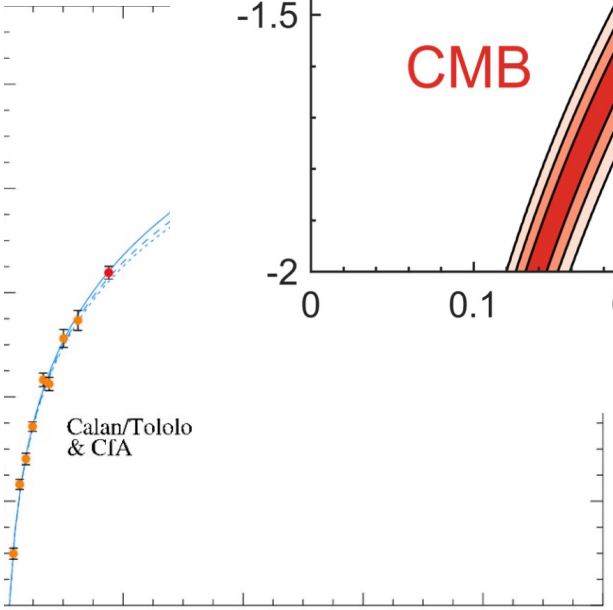
SCIENCE IS ABOUT DECISION

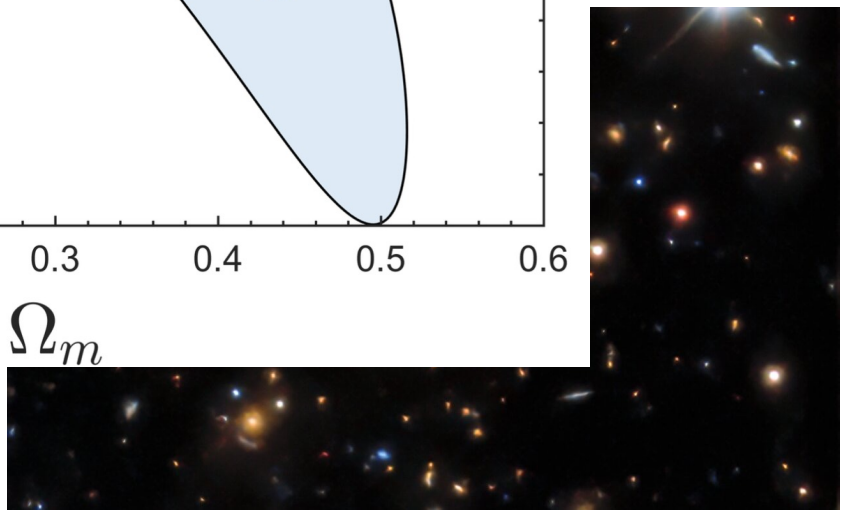
- *star vs galaxy ?*




12h
16h
8h
4h
0.10
0.20
0.30
0.40
Redshift z


BAO
Combined
SN Ia
CMB
-0.5
-1
-1.5
-2
0
0.1
0.2
0.3
0.4
0.5
0.6
w
Ω_m


Calan/Tololo
& CfA



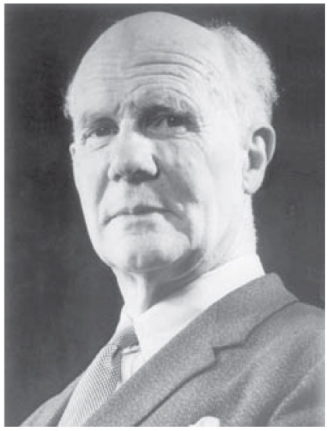
ASTRONOMERS CANNOT AVOID STATISTICS

- ***Data Error*** *(range):*
  - *How can data be used best? Or at all?*
  - *Correlation, testing the hypothesis, model fitting; how do we proceed?*
- ***Incomplete samples:***
  - *samples from an experiment cannot be re-run (cosmology)*
  - *upper limits*
- ***We must decide:***
  - *The decision process need some methodology*
  - *no matter how good the experiment.*

BAYESIAN VS. FREQUENTIST

Thomas Bayes
(1702–1761)

*What's **probability**?*
*How to handle **uncertainty?***
*How to incorporate **prior knowledge?***



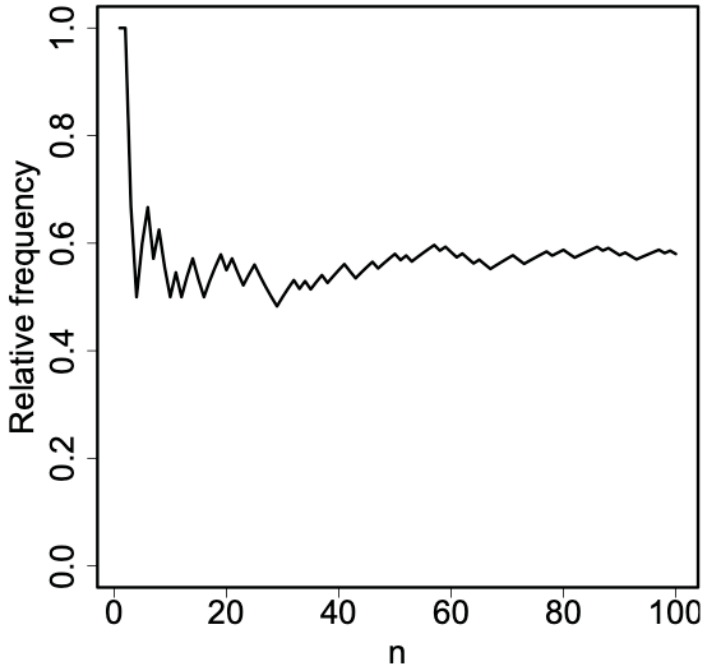
Jerzy Neyman, Egon Pearson and Ronald Fischer.

FREQUENTISTS

- *Consider experiments with a random component.*
- ***Probabilities:*** *the relative frequency of an event, A, is defined as*

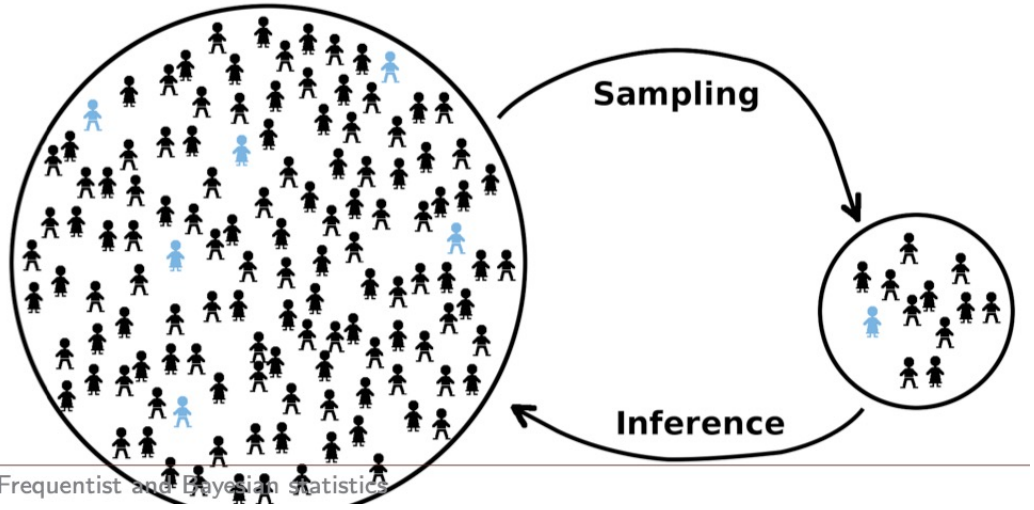
$$P(A) = \frac{\#\text{ of outcomes consistent with A}}{\#\text{ of experiments}}$$

- *The **probability** of event A is the **limiting** relative frequency*



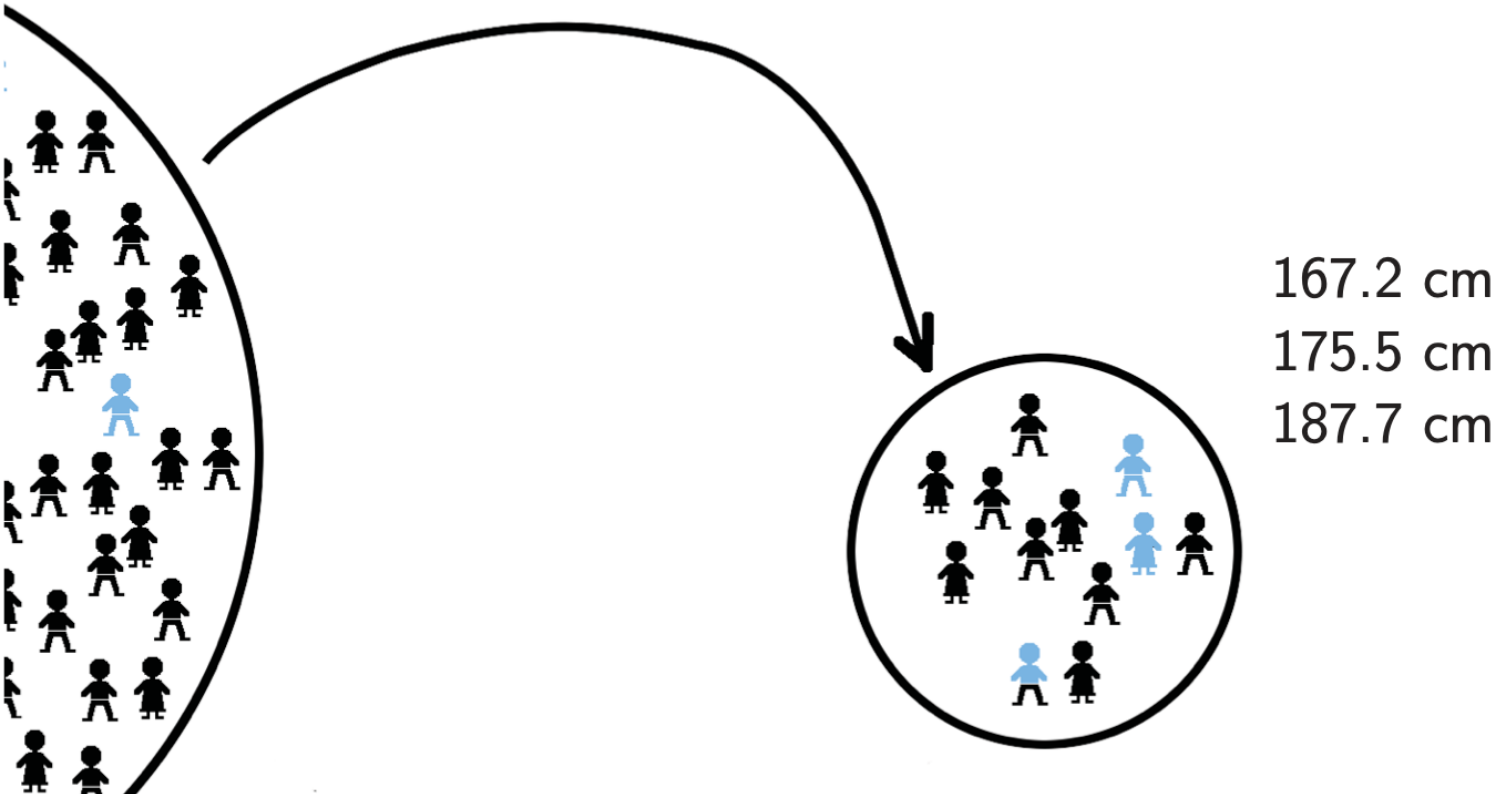
FREQUENTISTS

- *This definition **restricts** the things we can add probabilities to:*
  - *What is the probability of there being life on Mars 100 billion years ago?*
  - *not a repeatable experiment*
- *Assumption: there is an unknown but fixed underlying parameter, θ, for a population (e.g., the mean height of men).*
- *Random variation:*
  - *environmental factors, measurement errors, …*



THE META-EXPERIMENT IDEA

- *meta-experiments: consider the current dataset as a single realization from all possible datasets.*



THE META-EXPERIMENT IDEA

- *For example: a population mean is real but unknown, and unknowable*
  - *can only be estimated from the data.*
- *From the distribution for the sample mean, constructs a confidence interval, centered at the sample mean.*
  - *Regardless of the location of true mean;*
  - *Can't say there's a 95% probability that the true mean is in this interval, because it's either already in, or it's not.*
  - *the true mean is a fixed value, which doesn't have a distribution.*
  - *The sample mean does have a distribution!*
  - ***"95% of similar intervals would contain the true mean if each interval were constructed from a different random sample like this one."***

BAYESIANS

- *A numerical formalization of our degree of belief:*
  - *personal belief* → *the **prior:***
    - *No fixed values for parameters but a distribution.*
    - *All distributions are subjective: Yours is as good as mine*
- *Treat the parameters (e.g. mean) as random var.*
  - *mean : the mean of my distribution*
- *Only the data are real:*
  - *The population mean is an abstraction*
    - *Exist only conceptually*
  - *some values are more believable than others based on the data and their prior beliefs.*

BAYESIANS

- ***Credibility interval:***
  - *Posterior: centered near the sample mean, tempered by "prior".*
- *Bayesian can say what the frequentist cannot:*
  - *"There is a 95% probability (degree of believability) that this interval contains the mean."*

	Advantages	Disadvantages
Frequentist	Objective	Confidence intervals (not quite the desired)
	Calculations	
Bayesian	Credibility intervals (usually the desired)	Subjective
	Complex models	Calculations

IN SUMMARY

- *A frequentist is a person whose long-run ambition is to be wrong 5% of the time.*
- *A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.*

A frequentist uses impeccable logic to answer the wrong question, while a Bayesian answers the right question by making assumptions that nobody can fully believe in.

P. G. Hamer

EXAMPLE

- *Before 1987, 4 supernovae had been recorded in 10 centuries. What, before 1987, was the probability of a bright supernova happening in the 20th century?*
- *Three possible answers:*
  - *Probability is meaningless in this context.*
    - *Supernovae are physically determined events, not random.*
    - *From this God's-eye viewpoint, probability is indeed meaningless; 'God does not play dice…'*
  - *Frequentist: our best estimate of the probability is $4/10$, although it is obviously not very well determined.*
  - *Bayesian: a-priori assignment:*
    - *Modeling: stellar mass function, stellar birth rate, metallicity etc.*

PROBABILITY

中山大学

王 鑫

PROBABILITY

- *Measure of belief (Cox, 1946):*
  - *A, B and C are three events*
  - *wish to have some measure of how strongly we think each is likely to happen*
  - *apply the rule: if A is more likely than B, and B is more likely than C, then A is more likely than C.*
- *Axioms of probability (Kolmogorov)*
  - *any random event A has a probability $P(A)$ between 0 and 1;*
  - *the sure event has $P(A) = 1$;*
  - *If A and B are exclusive events, then $P(A \text{ or } B) = P(A) + P(B)$;*

CONDITIONALITY AND INDEPENDENCE

- *Two events A and B are said to be independent:*

$$P(A \text{ and } B) = P(A)P(B)$$

  - *Probability of one is unaffected by what we may know about the other.*

- *conditional probability:*

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

- *Marginalization:*

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

- *To get rid of these 'nuisance parameters' B_i.*

BAYES' THEOREM

- Bayes' theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- Particularly, θ is theory, d is data:

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}$$

  - posterior probability: $P(\theta|d)$
  - prior probability: $P(\theta)$, state of belief before the experiment
  - Likelihood function: $P(d|\theta) = \mathcal{L}(\theta), \mathcal{L}(d|\theta)$
  - normalizing factor: $P(d)$

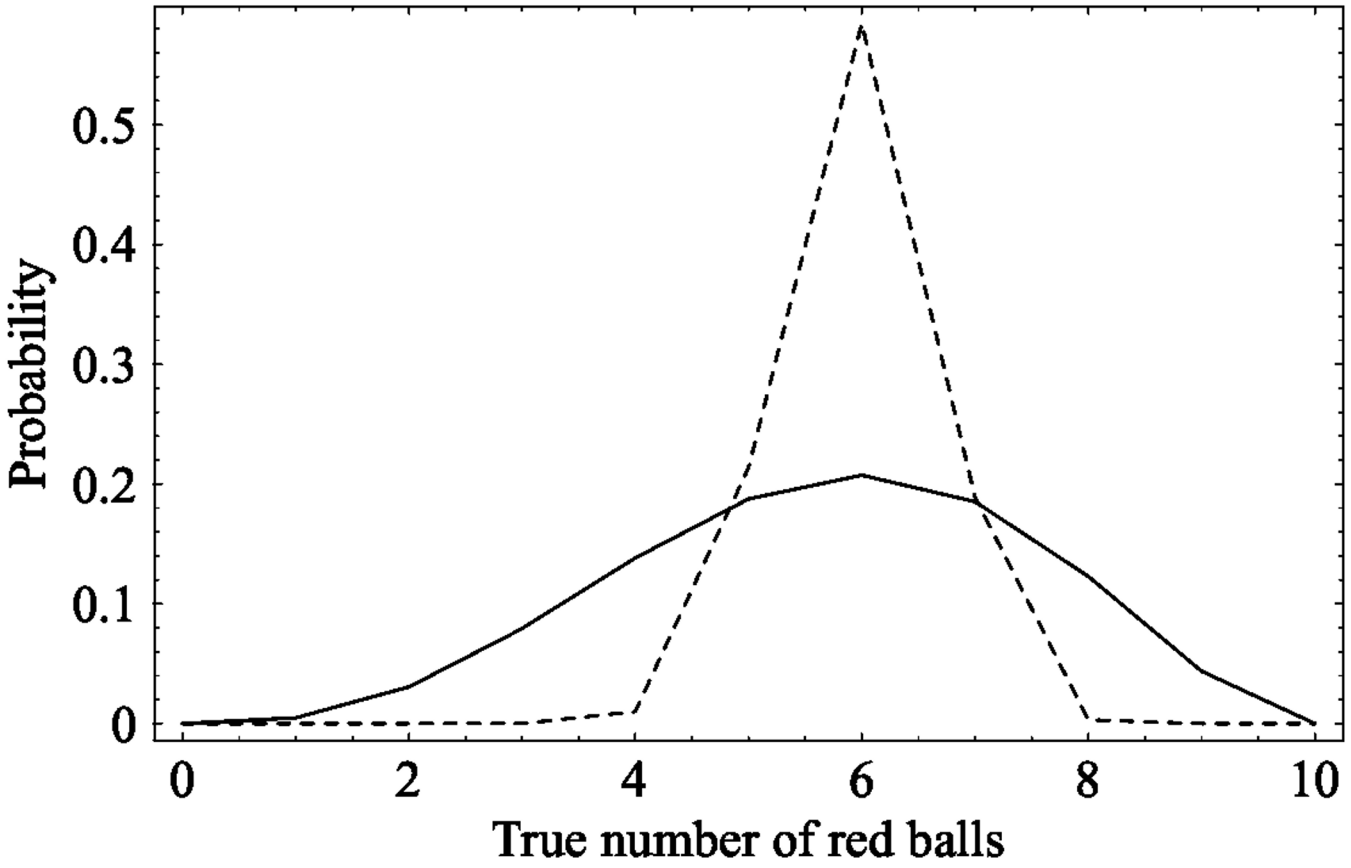
EXAMPLE

- There are N red balls and M blue balls in a jar, and $N + M = 10$. We draw $T = 3$ times (putting the balls back after drawing them) and $R = 2$ red balls. How many red balls are there in the jar?
- Model (hypothesis) probability of red ball: $N/(N + M)$
- The likelihood (probability of getting R red balls):

$$\binom{T}{R}\left(\frac{N}{N+M}\right)^{R}\left(\frac{M}{N+M}\right)^{T-R}$$

- Prior (uncertain):
  - uniformly likely between 0 and $N + M$

EXAMPLE



The probability distribution of the number of red balls, for five drawings (solid curve) and 50 drawings (dashed curve).

BINOMIAL DISTRIBUTION

- *2 outcomes: 'success' or 'failure':*
  - *each successive trials are independent, with probability ρ*
- *the chance of n successes in N trials:*

$$P(n) = \binom{N}{n} \rho^n (1-\rho)^{N-n}$$

- *The mean:*

$$\sum_{n=0}^{N} n\, P(n) = N\rho$$

- *The variance (mean square value):*

$$\sum_{n=0}^{N} (n - N\rho)^2 P(n) = N\rho(1-\rho)$$

EXAMPLE

- *Before 1987, 4 supernovae had been recorded in 10 centuries. What, before 1987, was the probability of a bright supernova happening in the 20th century?*
- *Define supernova rate per century ρ, posterior:*

$$P(\rho|data) \propto \binom{10}{4} \rho^4 (1-\rho)^6 \times (prior\ on\ \rho)$$

- *Take prior to be uniform, and normalize:*

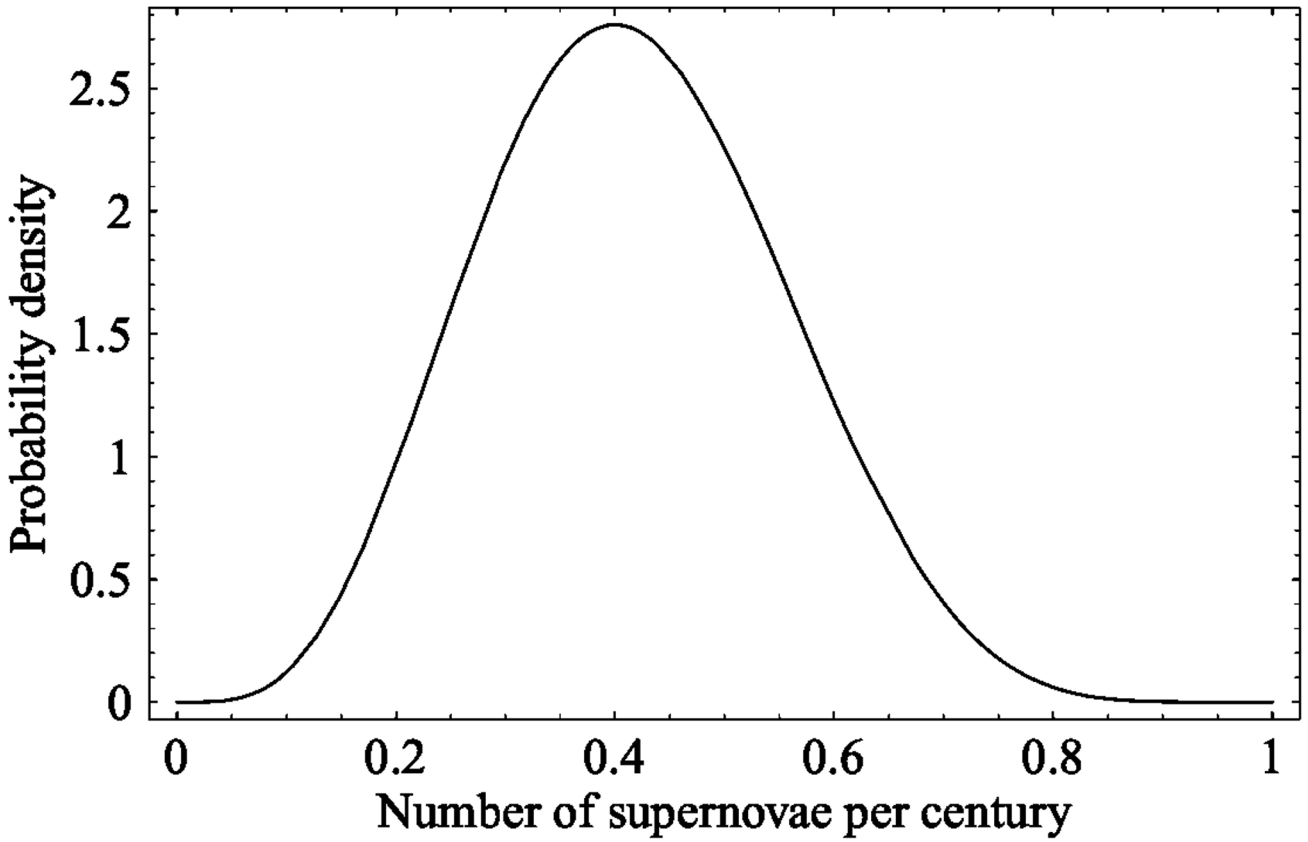
$$\int_0^1 P(\rho|data) d\rho = 1$$

- *for n supernovae in m centuries, the distribution is:*

$$P(\rho|data) = \rho^n (1-\rho)^{m-n} / B[n+1, m-n+1]$$

- *Here $B[n,m]$ is beta function.*

EXAMPLE



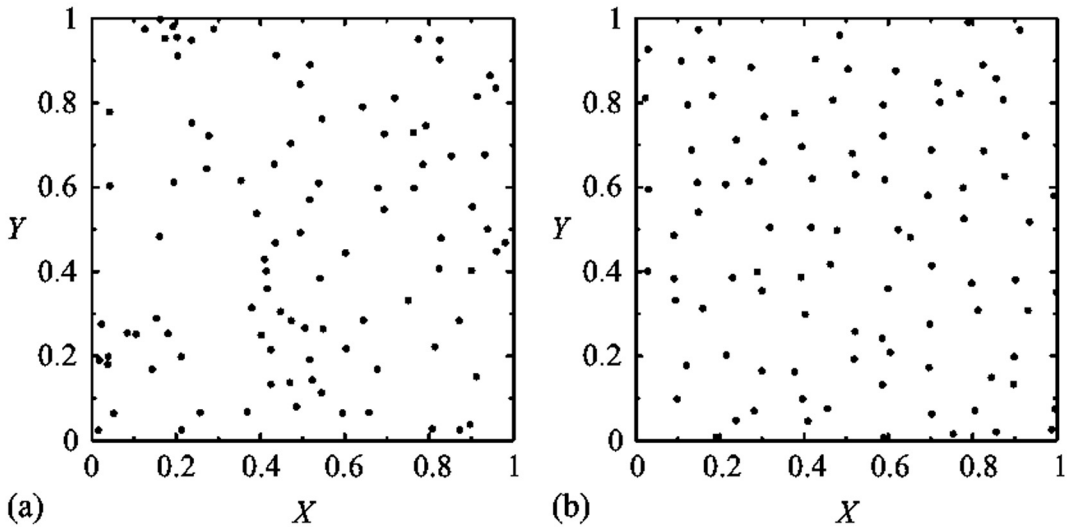
The posterior probability distribution for ρ, given that we have four supernovae in 10 centuries.

POISSON DISTRIBUTION

- *derives from the binomial in the limiting case of very rare (independent) events and a large number of trials:*
- *Binomial:* $\rho \to 0,\ N\rho \to \mu =$ *finite value.*
- *The probability of* n *events in a given interval, with the expectation of* μ *events in the same interval*

$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

- *The variance of the Poisson distribution is also* μ.



EXAMPLE

- *Poisson statistics govern the number of photons arriving during an integration.*
- *The probability of a photon arriving in a fixed interval of time is (often) small. The arrivals of successive photons are independent.*
- *Integration over time* t *of photons arriving at a rate* λ.
- *Mean photons:*

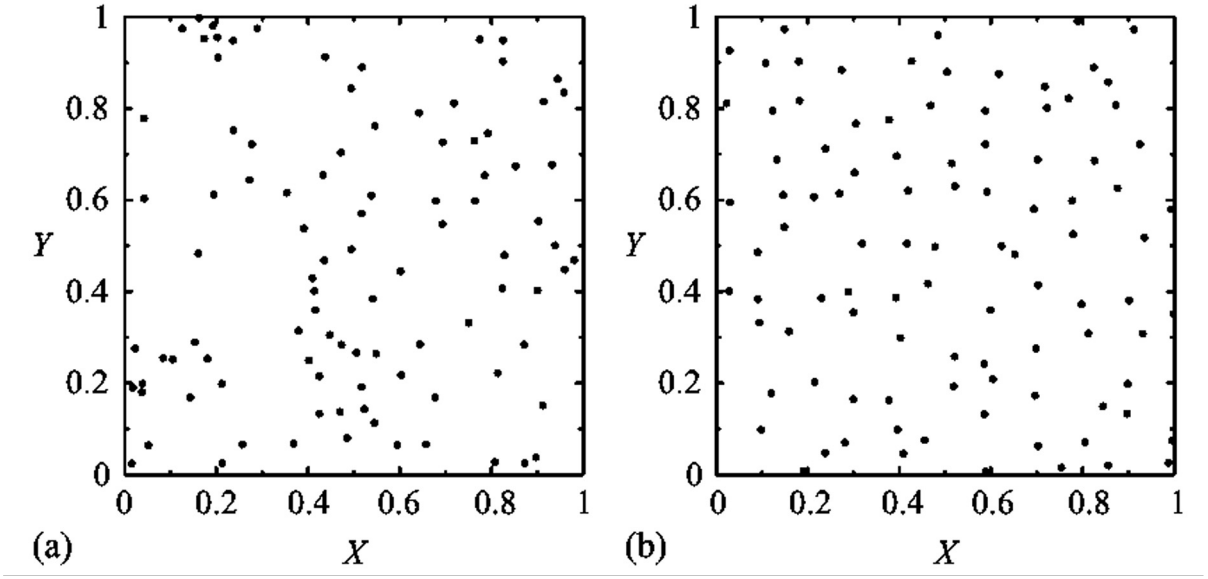
$$\mu = \lambda t$$

- *the fluctuation on this number will be*

$$\sigma = \sqrt{\mu}$$

EXAMPLE

- *In large-scale structure, the galaxy distribution is discrete;*
- *In a voxel of volume V, $\langle N \rangle$ average galaxies;*
- *The variance $\sigma_N^2 = \langle N \rangle = nV$;*
- *"shot noise" contribution of the galaxy distribution: $1/n$*



PROBABILITY DENSITY FUNCTION

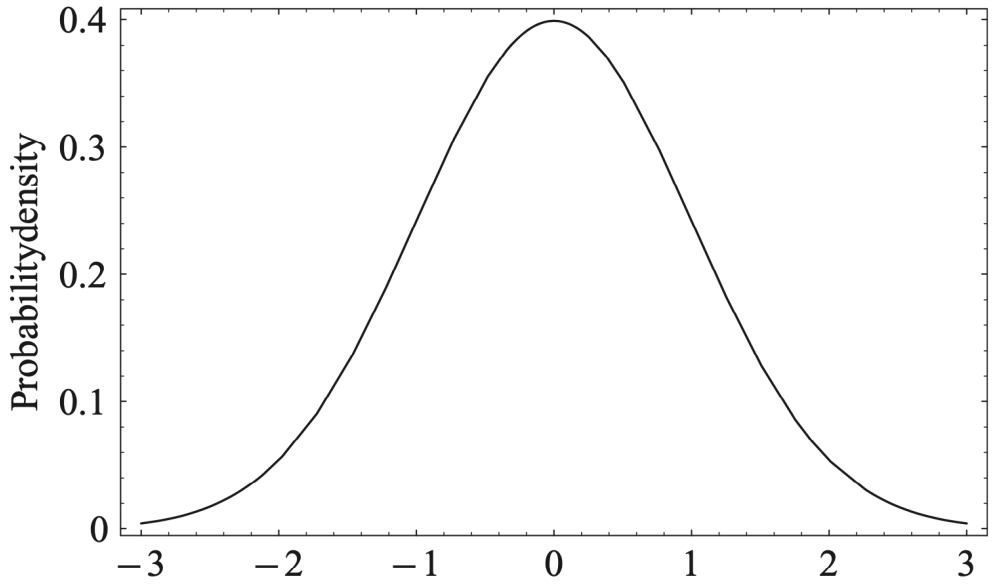
- *if x is a continuous random variable, then $f(x)$ is its probability density function (PDF), when:*
  - $P(a < x < b) = \int_a^b f(x)dx$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
  - $f(x)$ *is a single-valued non-negative number for all real x*
- *Cumulative probability distribution function:*

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

GAUSSIAN DISTRIBUTION

- *Large-N limit of both binomial and Poisson distributions:*

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



the area between 1σ is 0.68; between 2σ is 0.95; and between 3σ is 0.997.

CENTRAL LIMIT THEOREM:

- *Form averages* M_n

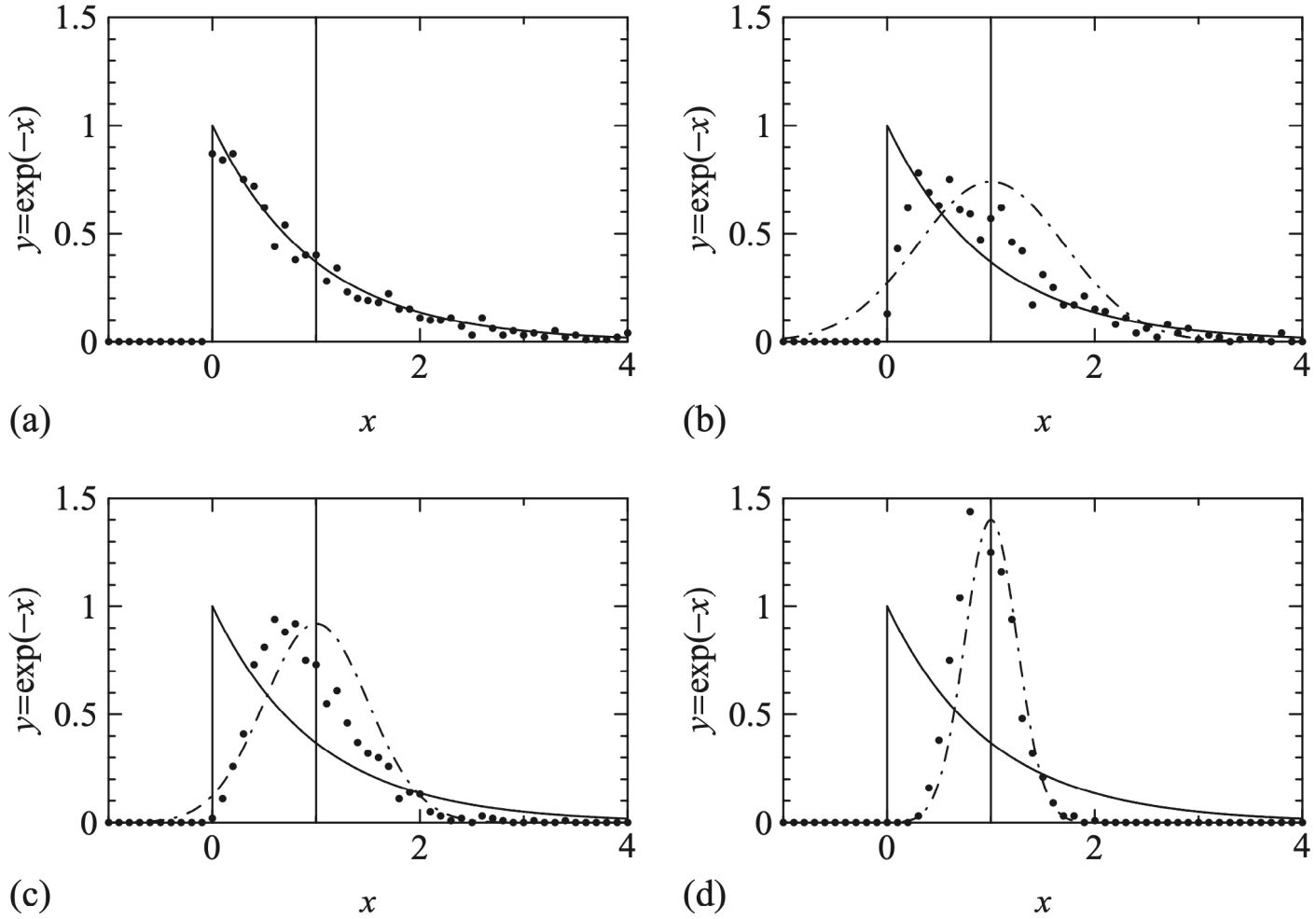
$$M_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *from repeatedly drawing* n *samples from a population* x_i *with finite mean* μ*, variance* σ^2*, then*

$$\lim_{n\to\infty}\left[\frac{M_n - \mu}{\sigma/\sqrt{n}}\right] \to N(0,1)$$

- *Averaging of large number of samples will produce Gaussian, no matter what the shape of the distribution from which the sample is drawn.*
  - *errors on averaged samples will always look 'Gaussian'*

CENTRAL LIMIT THEOREM:



An indication of the power of the central limit theorem. The panels show successive amounts of 'integration': in (a), a single value has been drawn; in (b), 200 values have been taken from an average of two values; (c), 200 values from an average of four; (d), 200 values from an average of 16.

STATISTICS

STATISTICS

- *Statistic: some function of the data alone*
- *Examples:*
  - *the location of the data:*
    - *Average:* $\bar{X} = 1/N \sum_{i=1}^{N} X_i$
    - *Median:* $X_{med} = X_j \ (j = N/2 + 0.5, N\ is\ odd; N/2\ N\ is\ even)$
    - *Mode:* X_{mode} *= value of* X_i *occurring most frequently, peak location in the histogram*
  - *the scale or amount of scatter in the data*
    - *Mean deviation:* $\overline{\Delta X} = 1/N \sum_{i=1}^{N} |X_i - X_{med}|$
    - *Mean square deviation:* $S^2 = 1/N \sum_{i=1}^{N} |X_i - \bar{X}|^2$
    - *Root-mean-square deviation:* $rms = S$

STATISTICS

- *some function f of a random variable x, with distribution function g, the expectation value:*

$$E[f(x)] = \langle f(x) \rangle = \int f(x) g(x) dx$$

- *With large # of experiments, the average of $\bar{X}$ will converge to the true mean value:*

$$E[x] = \langle x \rangle = \int x \; g(x) dx$$

- *Similarly, the statistics S^2 will converge to true variance*

$$\mathrm{var}[(x-\mu)^2] = E[(x-\mu)^2] = \int (x-\mu)^2 \; g(x) dx$$

- *n-th central moments*

$$\mu_n = \int (x-\mu)^n \; g(x) dx$$

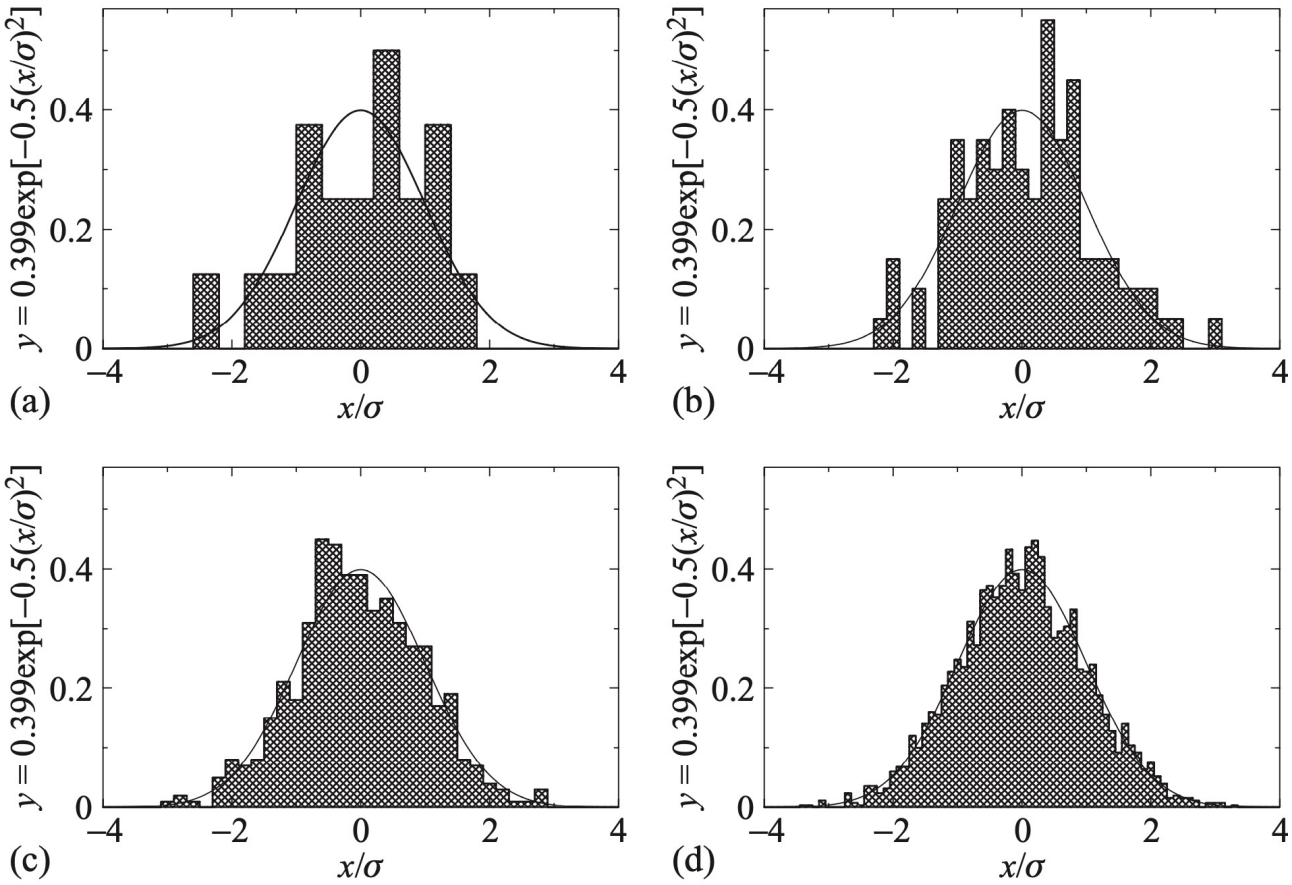
REQUIREMENTS FOR STATISTICS

- *unbiased*
  - *e.g. Gaussian distribution:*
    - $\bar{X} = 1/N \sum_{i=1}^{N} X_i$ *is indeed an unbiased estimate of mean* μ*;*
    - *unbiased estimation of the population variance* σ^2 *(sample variance):*

$$\sigma_s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

    - *differs from the expectation value of* S^2 *by the factor* $N/(N-1)$
    - $\bar{X}$*: yields a minimum value from the sum of the squares of the deviations, thus a low estimate of the variance.*

REQUIREMENTS FOR STATISTICS



x_i drawn at random from a Gaussian distribution of $\sigma = 1$: (a) 20 values, (b) 100 values, (c) 500 values, (d) 2500 values. The average values of x_i are 0.003, 0.080, 0.032 and 0.005; the median values 0.121, 0.058, 0.069 and 0.003; and the rms values 0.968, 1.017, 0.986 and 1.001. Solid curves represent Gaussians of unit area and standard deviation.

REQUIREMENTS FOR STATISTICS

- *consistent:*

$$\lim_{n\to\infty} \text{estimator} \to \text{true value}$$

  - rms *is a consistent measure of the standard deviation of a Gaussian distribution for large N;*
  - *but biased for small N.*

- *closeness*
  - *smallest possible deviation from the truth*

- *robust*
  - *robust against outliers*
  - *e.g. median is far more robust than average*

RANDOM VS. SYSTEMATIC

- *Variance on the average:*

$$S_m^2 = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right)^2\right]$$

$$\rightarrow S_m^2 = \frac{\sigma^2}{N} + \frac{1}{N^2}\sum_{i\neq j} E\left[(X_i - \mu)(X_j - \mu)\right]$$

- *Neglecting the second term, the error on the average scale with* $1/\sqrt{N}$.
- *Second term contains the covariance:*

$$\mathrm{cov}\left[X_i, X_j\right] = E\left[(X_i - \mu)(X_j - \mu)\right]$$

- *Describes the correlation between* x_i *and* x_j.

RANDOM VS. SYSTEMATIC

- For independent measurements (e.g. photometric measurements of some objects):

$$\mathrm{cov}[X_i, X_j] = \int (x_i - \mu_i) g(x_i|\cdots) dx_i \int (x_j - \mu_j) g(x_j|\cdots) dx_j = 0$$

- Here $g(x|some\ parameters)$ describe the PDF of X.
- Random vs. systematics:
  - random errors decrease with larger N ($1/\sqrt{N}$ or slower);
  - Systematic errors persist no matter how much data are collected
    - can only be reduced by better understanding the experiments

ERROR PROPAGATION

- *measure variables* $x, y, z \cdots$ *with independent errors* $\delta X, \delta Y, \delta Z \cdots$
- *interested in some function* $f(x, y, z, \cdots)$*, the error on* f

$$\delta f = \frac{\partial f}{\partial x}\Big|_{x=X} \delta X + \frac{\partial f}{\partial y}\Big|_{y=Y} \delta Y + \frac{\partial f}{\partial z}\Big|_{z=Z} \delta Z + \cdots$$

- *Assuming independent error:*

$$\mathrm{var}[f] = \left(\frac{\partial f}{\partial x}\right)^2\Big|_{x=X} \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2\Big|_{y=Y} \sigma_y^2 + \left(\frac{\partial f}{\partial z}\right)^2\Big|_{z=Z} \sigma_z^2 + \cdots$$

ERROR PROPAGATION

- *So we can have variance inverse weighting:*
  - *Data with larger variance are down-weighted.*
- *e.g. the weighted mean*

$$\bar{X}_w = \sum_{j=1}^{n} w_j \bar{X}_j \Big/ \sum_{j=1}^{n} w_j$$

- *The weight* $w_j = 1/\sigma_j^2$
- *The variance of* $\bar{X}_w$ *is then*

$$\sigma_w^2 = 1 \Big/ \sum_{j=1}^{n} 1/\sigma_j^2$$

COMBINING DISTRIBUTIONS

- Assuming we have the measured x, with PDF g, and function $f(x)$
- The PDF $h(f)$ of f could be derived because the conservation of the probability:

$$h(f)df = g(x)dx$$

- Example:
- $g(x) = exp(-x)$ for positive x, and $f(x) = log\, x$.
- It's easy to see that the PDF of f:

$$h(f) = exp[-exp(f)]\, exp(f)$$

COMBINING DISTRIBUTIONS

- *Assuming x, y follow PDF of $g(x), g(y)$, the PDF $h(z)$ of their sum $z = x + y$*

$$h(z) = \int g(z - x)g(x)dx$$

- *Is therefore a convolution*
- *Similarly, for $z = xy$:*

$$h(z) = \int \frac{1}{|x|} g(x)g\left(\frac{z}{x}\right) dx$$

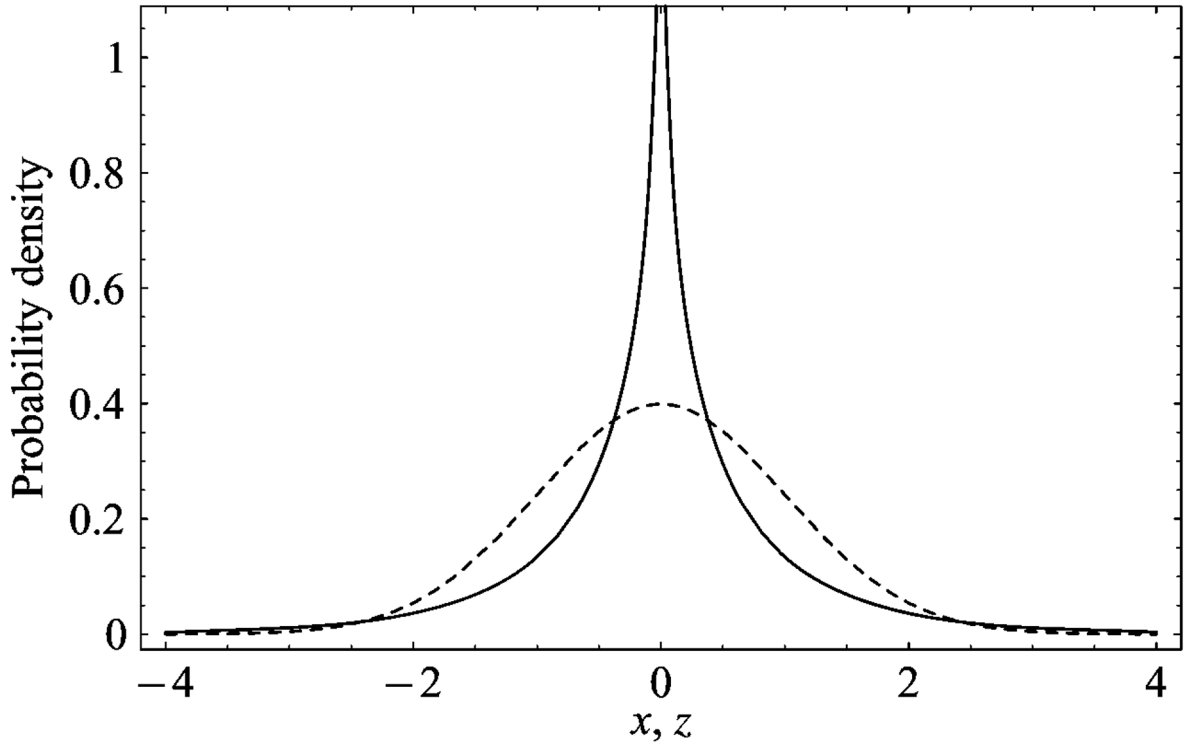
- *And for $z = x/y$:*

$$h(z) = \int |x|g(x)g(zx)dx$$

EXAMPLE

- *the product of two Gaussian variables of zero mean (e.g. in radio astronomy, visibility).*
- *The distribution of the product is (modified Bessel function)*

$$h(z) = \frac{2}{\pi\sigma^2} K_0\left[\frac{|z|}{\sigma^2}\right]$$

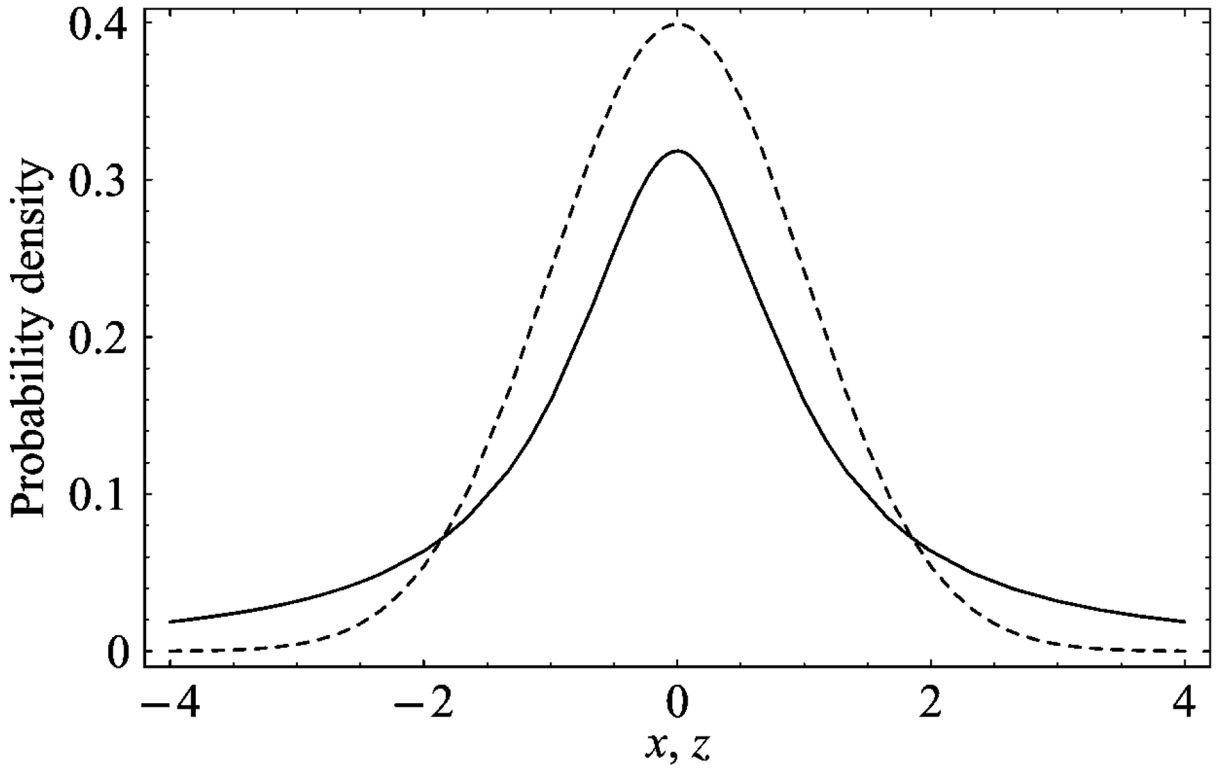


EXAMPLE

- *the ratio of two Gaussian variables of zero mean*

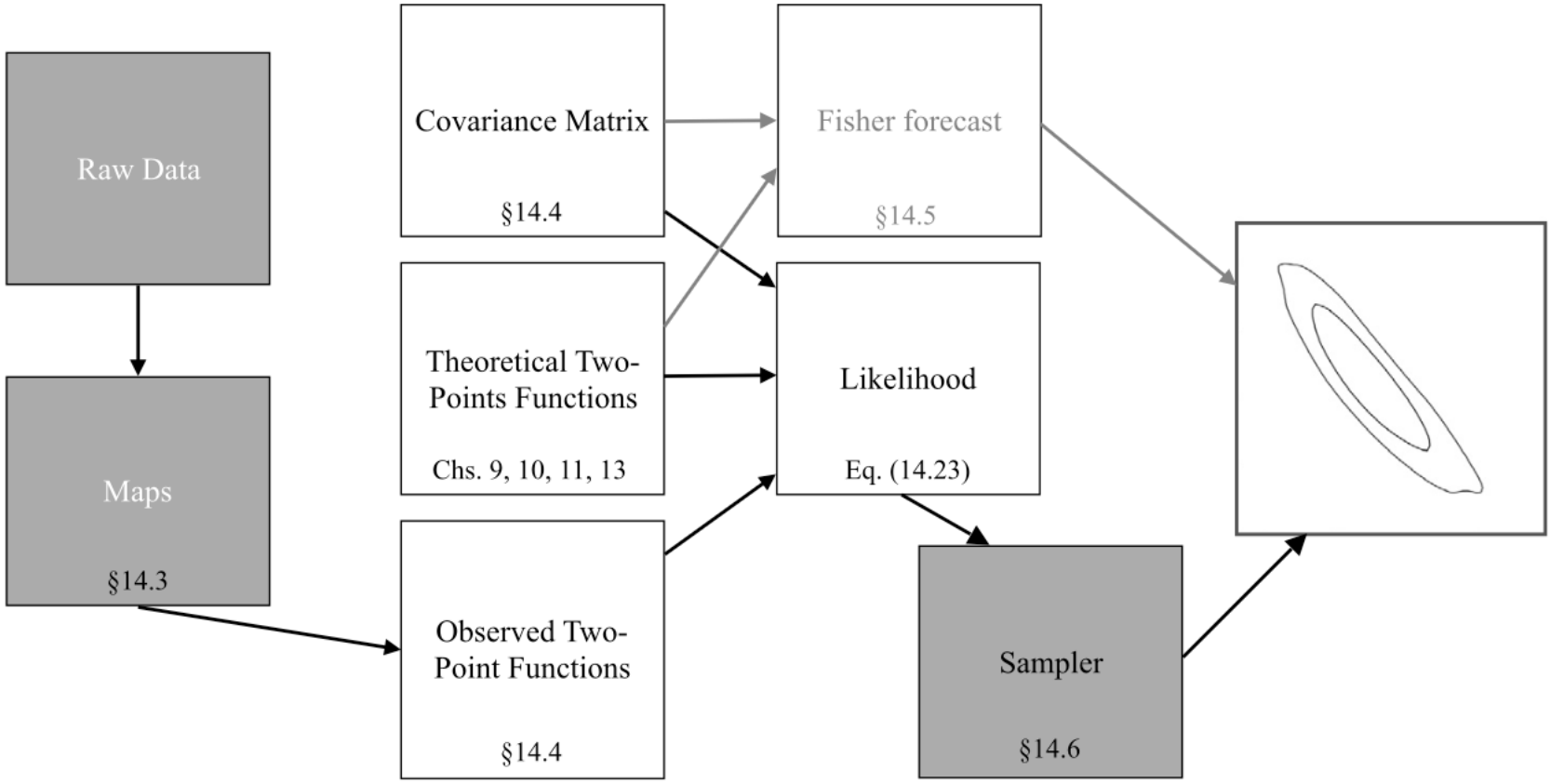
$$h(z) = \frac{1}{\pi}\frac{1}{1+z^2}$$

- *! Independent of* σ



LIKELIHOOD & INFERENCE

FROM RAW DATA TO PARAMETER



Modern cosmology, 2nd Edition

LIKELIHOOD

- *Suppose you want to weigh somebody:*

$$X_i = \mu(\vec{\alpha}) + n_i$$

LIKELIHOOD

- *Consider N data X_i, we estimate the statistics $(1/N)\sum_i X_i$*
- *It's a good estimator for $\mu(\vec{\alpha})$, where $\vec{\alpha} = (\alpha_1, \cdots, \alpha_n, \cdots)$ are unknown parameters (slopes, intercepts etc.)*
- *We believe that the measurement is scattered with Gaussian error around*

$$X_i = \mu(\vec{\alpha}) + n_i$$

- *with Gaussian distribution*

$$P(x|\vec{\alpha}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu(\vec{\alpha}))^2}{2\sigma^2}\right]$$

LIKELIHOOD

- From Bayes' theorem, the posterior probability distribution for the parameters $\vec{\alpha}$

$$P(\vec{\alpha}|X_i) \propto \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(X_i - \mu(\vec{\alpha})\right)^2}{2\sigma^2}\right] P(\vec{\alpha})$$

  - prior information $P(\vec{\alpha})$.
  - Here μ is "given", assuming everything depends on it being the correct model
  - Leads to the maximum likelihood method and method of least squares.
  - Easy to update models (posterior → prior)

MAXIMUM LIKELIHOOD

- ***Maximum likelihood*** (ML)
  - *derived by Bernoulli in 1776 and Gauss around 1821*
  - *worked out in detail by Fisher in 1922*
- *From the probability density function $f(x, \alpha)$, we'd like to estimate parameter α.*
- *Data $X_1, X_2, \cdots X_N$ independently drawn from f, the likelihood function is*

$$\mathcal{L}(X_1, X_2, \cdots X_N) = \prod^{N} f(X_i|\alpha)$$

- *assuming that the priors are 'diffuse', meaning that they change little over the peaked region of the likelihood function.*

MAXIMUM LIKELIHOOD

- Construct the '*best*' estimate of α: *peak* of $\mathcal{L}$
- Define *maximum likelihood estimator (MLE)* $\hat{\alpha}$:

$$\frac{\partial}{\partial \alpha} ln\, \mathcal{L}(\alpha) \Big|_{\alpha=\hat{\alpha}} = 0$$

- *MLE is a statistic:*
  - *It depends only on the data, not on parameters.*
- *This estimator would have some trouble if the priors are not diffuse*
  - *meaning they are having as strong an effect on our conclusions as the data*
- *minimum variance compared to any other estimate*
- *not always unbiased*

EXAMPLE

- *Measured Y_i at given independent variables X_i, with model*

$$y(a,b) = ax + b$$

- *Assuming Y_i have a Gaussian scatter, each term of likelihood:*

$$\mathcal{L}_i(y|(a,b)) = \exp\left[-\frac{\left(Y_i - (aX_i + b)\right)^2}{2\sigma^2}\right]$$

- *Maximizing the log-likelihood gives:*

$$\frac{\partial\, ln\, \mathcal{L}}{\partial a} = -2\sum_i X_i(Y_i - aX_i - b) = 0$$

$$\frac{\partial\, ln\, \mathcal{L}}{\partial b} = -2\sum_i (Y_i - aX_i - b) = 0;$$

- *This produce the ordinary least squares estimate:*

$$\hat{a} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - (\bar{X})^2}, \qquad \hat{b} = \bar{Y} - \hat{a}\bar{X}$$

EXAMPLE

- The source count of extragalactic radio sources:

$$N(>S) = kS^{-\gamma}$$

- N: the number of sources on a particular patch of sky with flux density greater than S.
- probability distribution:

$$P(S|\gamma) = dN/dS = \gamma k S^{-(\gamma+1)}$$

- Assuming observed M sources with flux densities S brighter than S_0, the normalization k:

$$\int_{S_0}^{\infty} P(S|\gamma)dS = 1 \rightarrow k = S_0^{\gamma}$$

EXAMPLE

- *So the likelihood function:*

$$\mathcal{L}(\gamma) = \prod_i^M P(S_i|\gamma) = \gamma^M S_0^{M\gamma} \prod_i^M S_i^{-(\gamma+1)}$$

- *So the log-likelihood:*

$$\ln \mathcal{L}(\gamma) = M \ln \gamma - \gamma \sum_i \ln \frac{S_i}{S_0} - \sum_i \ln S_i$$

- *Assuming observed M sources with flux densities S brighter than S_0*

$$\hat{\gamma} = M / \sum_i \ln \frac{S_i}{S_0}$$

EXAMPLE

- The intensity of neutrino burst after supernova decays exponentially after the core collapse. N neutrinos were detected with arrival times (in order) $T_1, T_2, \cdots$
- The probability of a neutrino arriving at time t is:

$$P(t) = \exp[-(t - t_0)]$$

- for $t > t_0$ and zero otherwise. To estimate parameter t_0,

$$ln\,\mathcal{L}(t_0) = Nt_0 - \sum_i T_i$$

- does not appear to have a maximum by derivative;
- But clearly $t_0 < T_1$, within the range, best estimator: $\hat{t}_0 = T_1$.

DEVIATIONS & FISHER MATRIX

- *MLE $\hat{\alpha}$ is distributed around true value;*
- *The covariance matrix of this distribution involves the curvature (Hessian matrix) of $\mathcal{L}$:*

$$\mathcal{H} = \begin{bmatrix} \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1^2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1 \partial \alpha_2} & \cdots \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_2 \partial \alpha_1} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_2^2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

- *This matrix depends on the data, taking the expectation value, we have the Fisher information matrix:*

$$F = E[\mathcal{H}]$$

DEVIATIONS & FISHER MATRIX

- the covariance matrix of the MLEs of the parameters:

$$C = F^{-1}$$

- Fisher matrix describes the width of the likelihood function, the scatter in the maximum-likelihood estimators
- The probability distribution of our N MLEs $\hat{\alpha}$ is then

$$P(\hat{\alpha}_1, \hat{\alpha}_2, \cdots)$$
$$= \frac{1}{\sqrt{(2\pi)^N |\, det\, C\,|}} exp\left[-\frac{1}{2}\left(\hat{\vec{\alpha}} - \vec{\alpha}\right) C^{-1} \left(\hat{\vec{\alpha}} - \vec{\alpha}\right)^T\right]$$

- Taking the expectation value is important, as otherwise the matrix would be different for each set of data

SIMPLE EXAMPLE

- A Gaussian of true mean μ and variance σ^2, if we have N data X_i, the log likelihood is

$$ln\,\mathcal{L} = -\frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2 - N\,ln\,\sigma$$

- And Fisher 'matrix':

$$F = -E\left(\frac{\partial^2 ln\,\mathcal{L}}{\partial \mu^2}\right) = \frac{N}{\sigma^2}$$

- Therefore, the variance on the estimate of the mean is

$$\frac{\sigma^2}{N}$$

- As expected~

EXAMPLE

- *In the source-count example, we have just one parameter, the variance on $\hat{\gamma}$ is then:*

$$-\frac{1}{E[\partial^2 \mathcal{L}(\gamma)/\partial \gamma^2]}$$

- *Which is*

$$\frac{\gamma^2}{M}$$

- *As long as the errors are the small, we can approximate*

$$\frac{\hat{\gamma}^2}{M}$$

BAYESIAN LIKELIHOOD ANALYSIS

- *From Bayes' theorem*

$$P(\vec{\alpha}|X_i) \propto \mathcal{L}(\vec{\alpha}|X_i)P(\vec{\alpha})$$

- *Two great strengths of the Bayesian approach:*
  - *deal with nuisance parameters via marginalization*
  - *the evidence or Bayes factor to choose between models*
- *asymptotic distribution of the likelihood function:*

$$\mathcal{L}(\vec{\alpha}|X_i) = \mathcal{L}(\hat{\vec{\alpha}}|X_i)\exp\left(-\frac{1}{2}(\hat{\vec{\alpha}}-\vec{\alpha})F(\hat{\vec{\alpha}}-\vec{\alpha})^T\right)$$

- *Here F is the Fisher information matrix.*
- *This is called the Laplace approximation.*

BAYES FACTOR

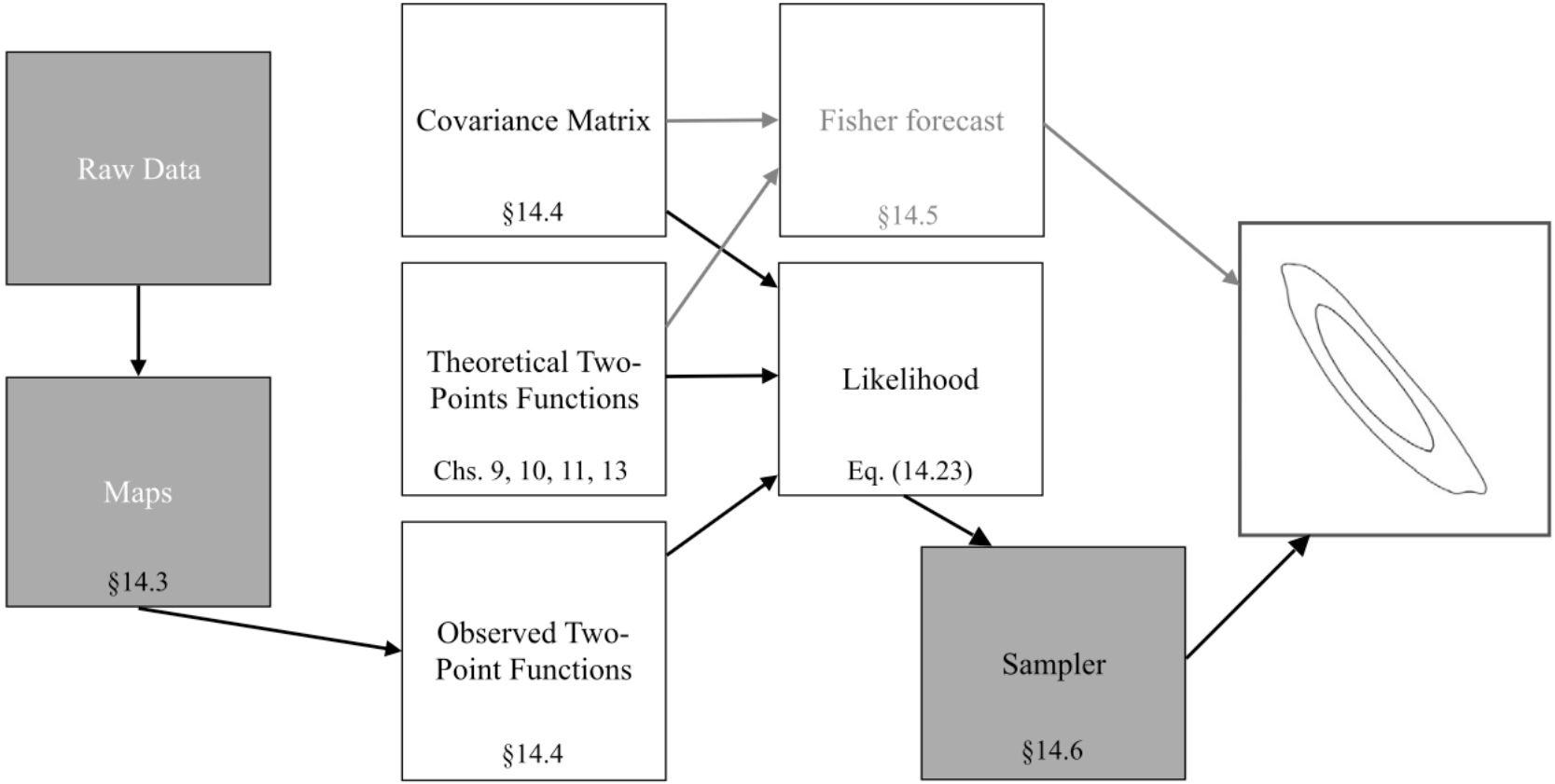
- Need to check the 'fit' of the two models:
  - choice of Model A or Model B
- Using the Bayes factor, to calculate the posterior odds on Model A, compared to Model B,

$$\mathcal{P} = \frac{\int_\alpha p_A \mathcal{L}(X_i|\alpha, A) P(\alpha|A)}{\int_\alpha p_B \mathcal{L}(X_i|\alpha, B) P(\alpha|B)}$$

- The Integration might be cumbersome, but it's worth the effort.

COSMOLOGICAL DATA ANALYSIS

FROM RAW DATA TO PARAMETER CONSTRAINTS



$$\ln \mathcal{L}(\lambda_\alpha) = -\frac{1}{2} \sum_{ll'} \left(\hat{C}(l) - C^{\text{theory}}(l, \lambda_\alpha) \right) \left(\text{Cov}^{-1} \right)_{ll'} \left(\hat{C}(l') - C^{\text{theory}}(l', \lambda_\alpha) \right)$$

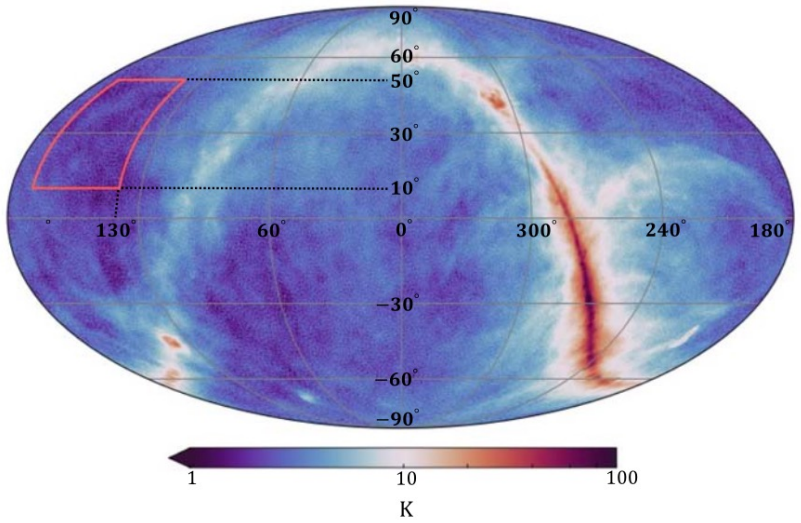
MAPMAKING

MAPMAKING

- *Assume the observed time ordered data*

$$d_t = P_{ti} s_i + \eta_t$$

- s_i *is sky map with pixel index of* i, η_t *is noise, and* P_{ti} *is the pointing matrix.*



- *To extract the signal from data, consider the likelihood*

$$\chi^2 = -2 \ln \mathcal{L}(\{d_t\}|\{s_k\})$$

$$= \sum_{tt'kl} (d_t - P_{tk} s_k)(N^{-1})_{tt'} (d_{t'} - P_{t'l} s_l)$$

MAPMAKING

- *So the MLE is*

$$\frac{\partial \chi^2}{\partial s_i} = -2 \sum_{tt'j} P_{ti} (N^{-1})_{tt'} (d_{t'} - P_{t'l} s_l) = 0$$

- *Which leads to*

$$\hat{s}_i = \sum_{tt'j} (C_N)_{ij} P_{tj} (N^{-1})_{tt'} d_{t'}$$

$$(C_N^{-1})_{ij} = \sum_{tt'} P_{ti} \, (N^{-1})_{tt'} P_{t'j}$$

$$\rightarrow \hat{s} = C_N P^T N^{-1} d$$

TWO-PIONT FUNCTIONS

CMB POWER SPECTRUM

- *Transfer the observed map into spherical harmonics*

$$a_{lm}^{\rm obs} = \int d\Omega\, Y_{lm}^*(\hat{\boldsymbol{n}})\Delta(\hat{\boldsymbol{n}})$$

- *the fractional temperature fluctuation* $\Delta = (T - T_0)/T_0$

$$\Delta(\hat{\boldsymbol{n}}) = \int d\Omega' \Theta(\hat{\boldsymbol{n}}')B(\hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}') + \eta(\hat{\boldsymbol{n}})$$

- *Combine two definitions, one has*

$$a_{lm}^{\rm obs} = \sum_{l'm'} B_{lm,l'm'} a_{l'm'} + \eta_{lm}$$

- *isotropic beam* $\rightarrow$ $a_{lm}^{\rm obs} = a_{lm} B_l + \eta_{lm}$

CMB POWER SPECTRUM

- *The likelihood function of*

$$\langle a_{lm} \rangle = 0; \qquad \langle a_{lm} a^*_{l'm'} \rangle = \delta_{ll'} \delta_{mm'} C(l)$$

$$\langle \eta_{lm} \eta^*_{l'm'} \rangle = N(l) \delta_{ll'} \delta_{mm'}$$

$$P\left(\{a_{lm}^{\rm obs}\}|C(l)\right) = \prod_{m=-l}^{l} \int da_{lm}\, P(a_{lm}^{\rm obs}|a_{lm}) P(a_{lm}|C(l))$$

$$P(a_{lm}^{\rm obs}|a_{lm}) = \frac{1}{\sqrt{2\pi N(l)}} \exp\left[-\frac{1}{2N(l)}|a_{lm}^{\rm obs} - B_l a_{lm}|^2\right]$$

- *The probability $P(a_{lm}|C(l))$ is Gaussian with mean zero and variance $C(l)$;*
- *Carry out the integral, we have*

$$\mathcal{L} \equiv P\left(\{a_{lm}^{\rm obs}\}|C(l)\right) = \left(2\pi\left[C(l)B_l^2 + N(l)\right]\right)^{-(2l+1)/2} \exp\left\{-\frac{1}{2}\sum_{m=-l}^{l} \frac{\left|a_{lm}^{\rm obs}\right|^2}{C(l)B_l^2 + N(l)}\right\}$$

CMB POWER SPECTRUM

- *The first derivative of the log of the likelihood*

$$\frac{d\ln\mathcal{L}}{dC(l)} = -\frac{(2l+1)B_l^2/2}{C(l)B_l^2 + N(l)} + \frac{1}{2}\sum_{m=-l}^{l}\frac{\left|a_{lm}^{\mathrm{obs}}\right|^2 B_l^2}{[C(l)B_l^2 + N(l)]^2}$$

- *Setting to zero, obtain the estimator for $C(l)$ is:*

$$\hat{C}(l) = B_l^{-2}\left(\frac{1}{2l+1}\sum_{m=-l}^{l}\left|a_{lm}^{\mathrm{obs}}\right|^2 - N(l)\right)$$

- *The error of this estimator:*

$$\mathrm{Var}\left[\hat{C}(l)\right] = \left\langle \hat{C}(l)^2 \right\rangle - C(l)^2$$

CMB POWER SPECTRUM

- *Expand the first term and use* $\langle |a_{lm}^{\rm obs}|^2 \rangle = C(l) B_l^2 + N(l)$

$$\left\langle B_l^{-4} \left(\frac{1}{2l+1} \sum_{m=-l}^{l} \left| a_{lm}^{\rm obs} \right|^2 - N(l) \right)^2 \right\rangle - C(l)^2 =$$

$$= \left\langle B_l^{-4} \left(\frac{1}{2l+1} \sum_{m=-l}^{l} \left| a_{lm}^{\rm obs} \right|^2 \right)^2 \right\rangle \quad \Rightarrow \quad = \frac{2l+3}{2l+1} \left[C(l) + N(l) B_l^{-2} \right]^2$$

$$- 2B_l^{-4} N(l) \left(C(l) B_l^2 + N(l) \right) + B_l^{-4} N(l)^2 - C(l)^2$$

- *So the error on estimator*

$$\sqrt{\mathrm{Var}\left[\hat{C}(l) \right]} = \sqrt{\frac{2}{2l+1}} \left[C(l) + N(l) B_l^{-2} \right]$$

$$\mathrm{Cov}_{ll'} = \frac{2}{2l+1} \left[C(l) + N(l) B_l^{-2} \right]^2 \delta_{ll'}$$

GALAXY POWER SPECTRUM

- *discrete Fourier transform of the galaxy density field:*

$$\delta_{\rm g}(\boldsymbol{k}) = L^{3/2} \sum_{i}^{K^3_{\rm grid}} \delta_{\rm g}(\boldsymbol{x}_i) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}_i}, \quad \text{where} \quad \boldsymbol{k} \in (n_x, n_y, n_z)\, k_F$$

- *The fundamental frequency* $k_F = 2\pi/L$
- *Similar to CMB* $C(l)$*, and setting* $B_l = 1$*, we have estimator*

$$\hat{P}_{\rm g}(k_\alpha) = \frac{1}{m_{k,\alpha}} \sum_{\boldsymbol{k}}^{||\boldsymbol{k}|-k_\alpha|<\Delta k/2} |\delta_{\rm g}(\boldsymbol{k})|^2 - P_N$$

- *The number of modes in each shell:*

$$m_{k,\alpha} = \frac{1}{2}\frac{4\pi k_\alpha^2 \Delta k}{k_F^3} = \frac{1}{4\pi^2} V k_\alpha^2 \Delta k$$

GALAXY POWER SPECTRUM

- *The covariance matrix:*

$$\mathrm{Cov}_{\alpha\beta} \equiv \left\langle \hat{P}_{\mathrm{g}}(k_\alpha)\hat{P}_{\mathrm{g}}(k_\beta)\right\rangle - \left\langle \hat{P}_{\mathrm{g}}(k_\alpha)\right\rangle \left\langle \hat{P}_{\mathrm{g}}(k_\beta)\right\rangle$$

$$= \frac{1}{m_{k,\alpha}} \sum_{\boldsymbol{k}}^{||\boldsymbol{k}|-k_\alpha|<\Delta k/2} \frac{1}{m_{k,\beta}} \sum_{\boldsymbol{k}'}^{||\boldsymbol{k}'|-k_\beta|<\Delta k/2} \left[\left\langle |\delta_{\mathrm{g}}(\boldsymbol{k})|^2 |\delta_{\mathrm{g}}(\boldsymbol{k}')|^2\right\rangle - \left\langle |\delta_{\mathrm{g}}(\boldsymbol{k})|^2\right\rangle \left\langle |\delta_{\mathrm{g}}(\boldsymbol{k}')|^2\right\rangle\right]$$

- *Under Gaussian assumption, this eventually leads to*

$$\mathrm{Cov}_{\alpha\beta} = \frac{2}{m_{k,\alpha}} \left[P_{\mathrm{g}}(k_\alpha) + P_N\right]^2 \delta_{\alpha\beta}$$

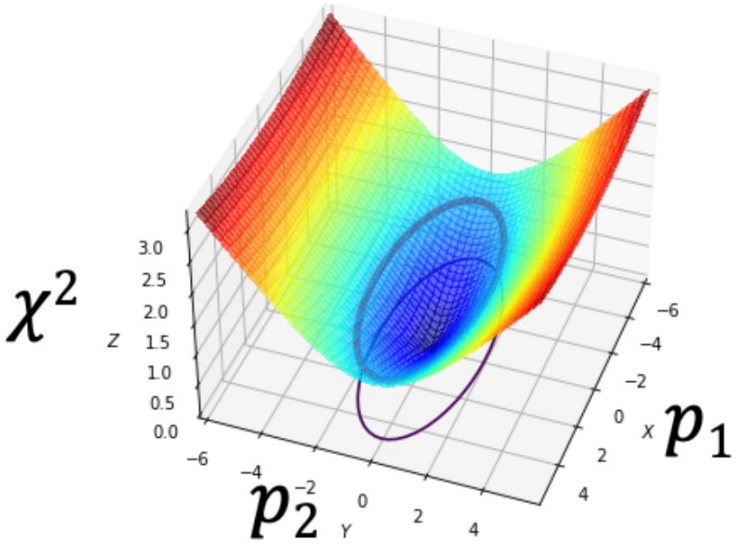
SAMPLING THE LIKELIHOOD FUNCTION

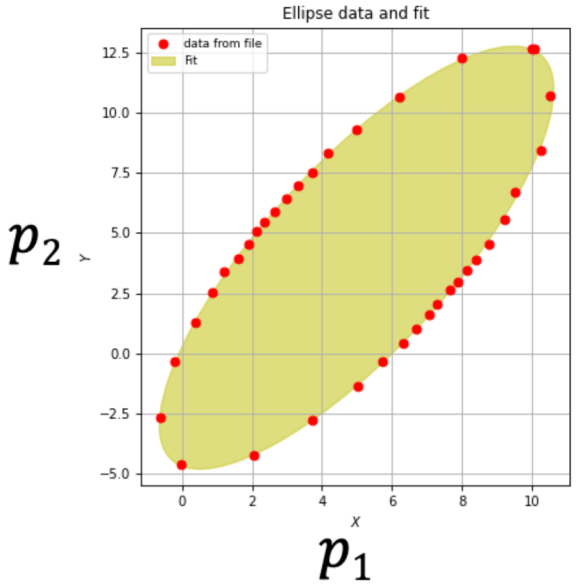
A DIRECT METHOD: GRID SCAN

- *Divide parameter space into small grids*
- *Find the minimum chi-square* $\chi^2_{\min}$
- *Find confidence levels by*

$$\Delta\chi^2 = \chi^2 - \chi^2_{\min}$$

- *Limitation:*
- *Time-consuming:* $t \sim n_{\text{grid}}^N$, N *is the number of parameters*
- *Marginalization is complicated!*





SAMPLING

- *Purposes:*
- *To generate samples* $\{\boldsymbol{x}^{(r)}\}_{r=1}^{R}$ *from a given probability distribution* $P(\boldsymbol{x})$
- *To estimate expectation of a function* $\phi(\boldsymbol{x})$ *under this distribution* $P(\boldsymbol{x})$

$$\Phi = \langle \phi(\boldsymbol{x}) \rangle = \int d^N x \, P(\boldsymbol{x}) \phi(\boldsymbol{x})$$

- *Therefore,*

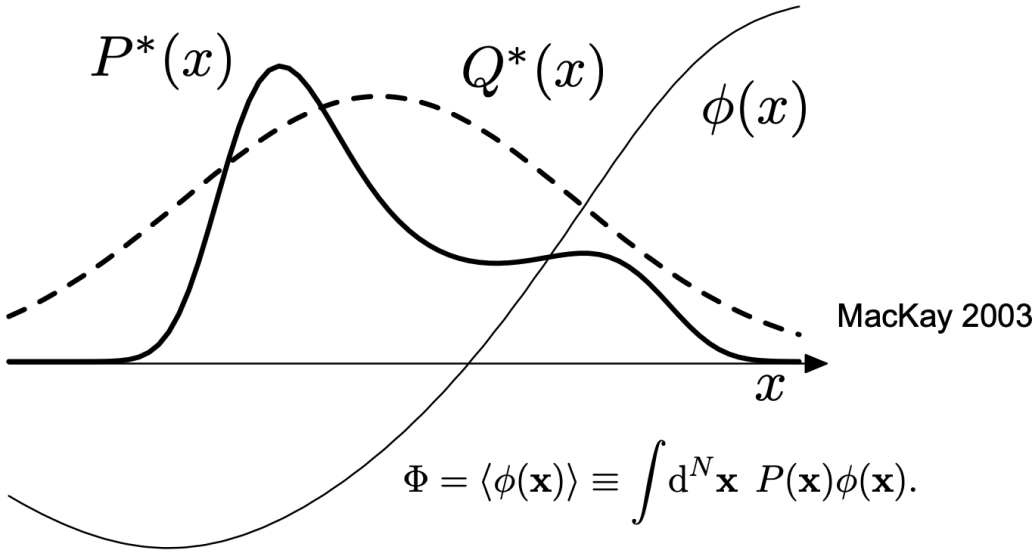
$$\Phi = \frac{1}{R} \sum_{r} \phi\left(\boldsymbol{x}^{(r)}\right)$$

SAMPLING METHODS

- *Monte Carlo (MC): Random sampling*
  - *Uniform sampling*
  - *Importance sampling*
  - *Rejection sampling*
- *Markov chain Monte Carlo (MCMC)*
  - *Metropolis-Hastings method*
  - *Gibbs sampling*
  - *Slice sampling*
  - *Hamiltonian Monte Carlo*
  - ……

IMPORTANCE SAMPLING

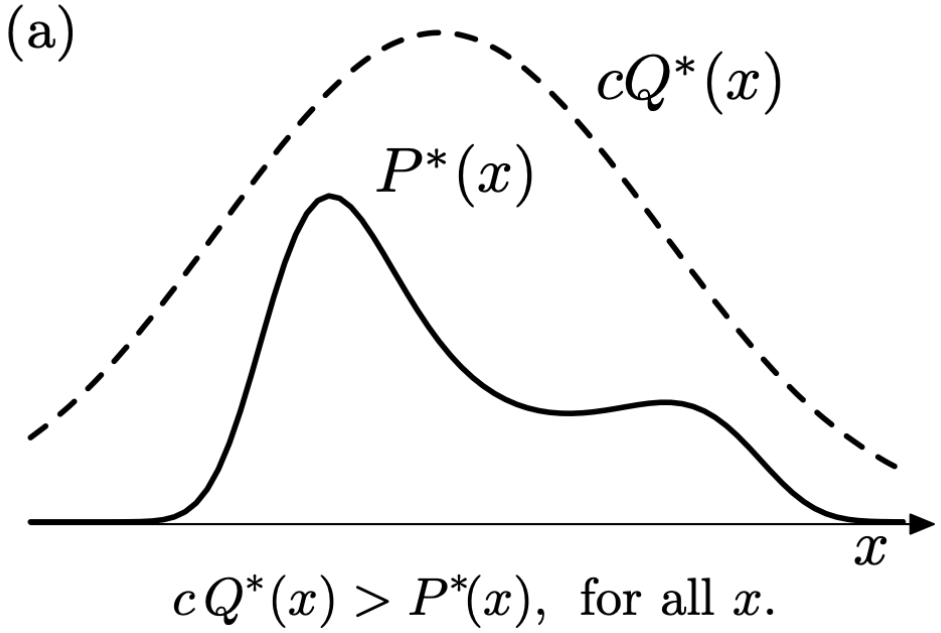
- *Generate samples from a simpler (wrong) distribution* $Q(x)$
- *Assign high probability to "important" values*
  - *adjusting the "importance" of each point by introducing a weight* $w_r = P^*(\boldsymbol{x}^{(r)})/Q^*(\boldsymbol{x}^{(r)})$
- *The expectation is* $\Phi = \sum_r w_r \phi(\boldsymbol{x}^{(r)}) \,/\, \sum_r w_r$



REJECTION SAMPLING

- Selecting a simpler proposal distribution $Q(x)$, and find a constant c such that $cQ(x) > P(x)$ for all x;
- Generating two random numbers:
  - Sample x from $Q(x)$;
  - Generate a uniformly distributed random variable u from the interval $[0, cQ(x)]$;
- If $u > P(x)$, x is rejected;
- else, it is accepted;

REJECTION SAMPLING



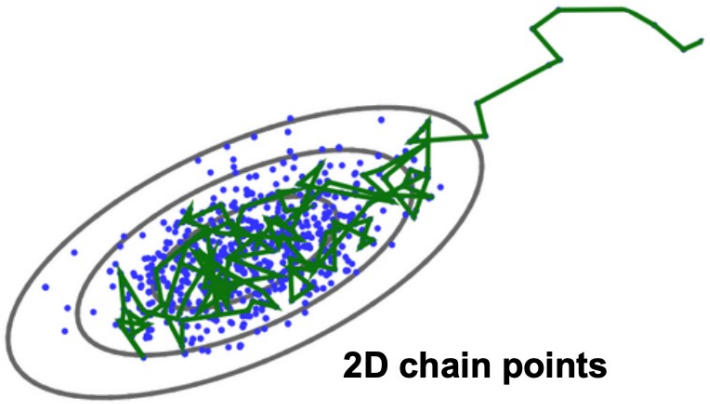
- *Work best if Q is a good approximation to P;*
- *otherwise acceptance rate will be too low;*
- *Not suitable in high-dimension (N parameters) case*

MacKay 2003

MARKOV CHAIN MONTE CARLO

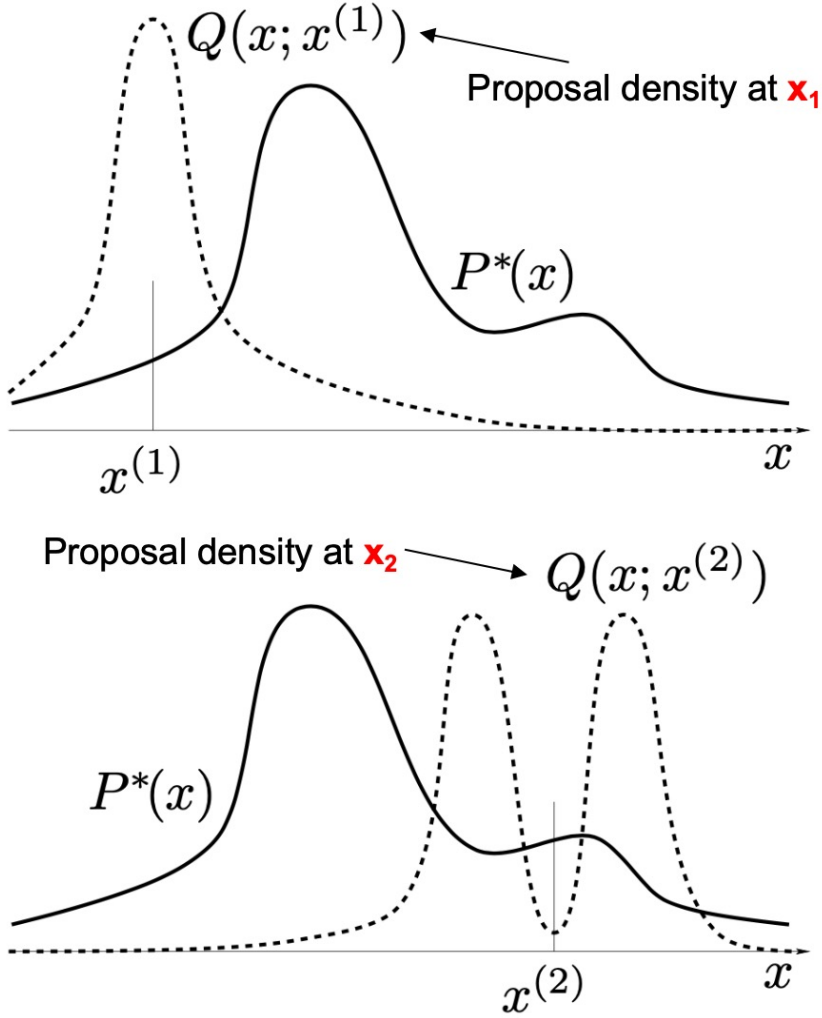
- *Based on simulation;*
- *Computational time cost is approximately linearly with the number of parameters* $t \sim N$*;*
- *Generating samples from the full posterior distribution, easy marginalization;*
- *Algorithm:*
  - *Metropolis-Hastings, Gibbs, Slice sampling, Hamiltonian MC, …*

MARKOV CHAIN MONTE CARLO



- *Markov Chain: a chain composed by a sequence of steps (or chain points), that the next 'step' in the sequence depends only upon the previous one;*
- *Monte Carlo: a computational algorithms which rely on random sampling, with the algorithm being guided by some rules designed to give the desired outcome*

METROPOLIS-HASTINGS METHOD



MacKay 2003

- *It makes use of a proposal density Q which depends on the current state x. Proposal density at x. Q can be any fixed density from which we can draw samples;*
- *First a new sample is proposed based on the previous sample, then the proposed sample is either added to the sequence or rejected depending on the value of the probability distribution at that point;*
- *It generally used for sampling from multi dimensional distributions, especially when the number of dimensions is high.*

METROPOLIS-HASTINGS METHOD

- *MH method obtains the MCMC chain points by applying the acceptance probability, which is defined as:*

$$\mathbf{a}(\theta_{\mathbf{n+1}}|\theta_{\mathbf{n}}) \;=\; \min\left\{\frac{\mathbf{p}(\theta_{\mathbf{n+1}}|\mathbf{d})\;\mathbf{q}(\theta_{\mathbf{n}}|\theta_{\mathbf{n+1}})}{\mathbf{p}(\theta_{\mathbf{n}}|\mathbf{d})\;\mathbf{q}(\theta_{\mathbf{n+1}}|\theta_{\mathbf{n}})}\;,\mathbf{1}\right\}$$

- θ_n *is the nth chain point (or a parameter set),* $\boldsymbol{d}$ *is the observational data,* $p(\theta|\boldsymbol{d})$ *is the posterior distribution,* $q(\theta_{n+1}|\theta_n)$*is the proposal density.*

METROPOLIS-HASTINGS METHOD

- Based on Bayes' theorem, posterior distribution can be estimated by $$\mathbf{p}(\theta|\mathbf{d}) = \frac{\mathcal{L}(\mathbf{d}|\theta)\,\mathbf{p}(\theta)}{\int \mathcal{L}(\mathbf{d}|\theta)\mathbf{p}(\theta)\mathbf{d}\theta}$$
- $p(\theta)$ is the prior probability, and $\mathcal{L}(\boldsymbol{d}|\theta_n)$ is the likelihood function.
- When assuming an uniform prior distribution, we have $p(\theta|\boldsymbol{d}) \sim \mathcal{L}(\boldsymbol{d}|\theta_n)$
- Besides, if assume that the proposal density follows the same Gaussian distribution for every chain point, we have $q(\theta_{n+1}|\theta_n) = q(\theta_n|\theta_{n+1})$, then we find:

$$\mathbf{a}(\theta_{\mathbf{n+1}}|\theta_{\mathbf{n}}) = \min\left\{\frac{\mathcal{L}(\mathbf{d}|\theta_{\mathbf{n+1}})}{\mathcal{L}(\mathbf{d}|\theta_{\mathbf{n}})}\,, 1\right\}$$

METROPOLIS-HASTINGS METHOD

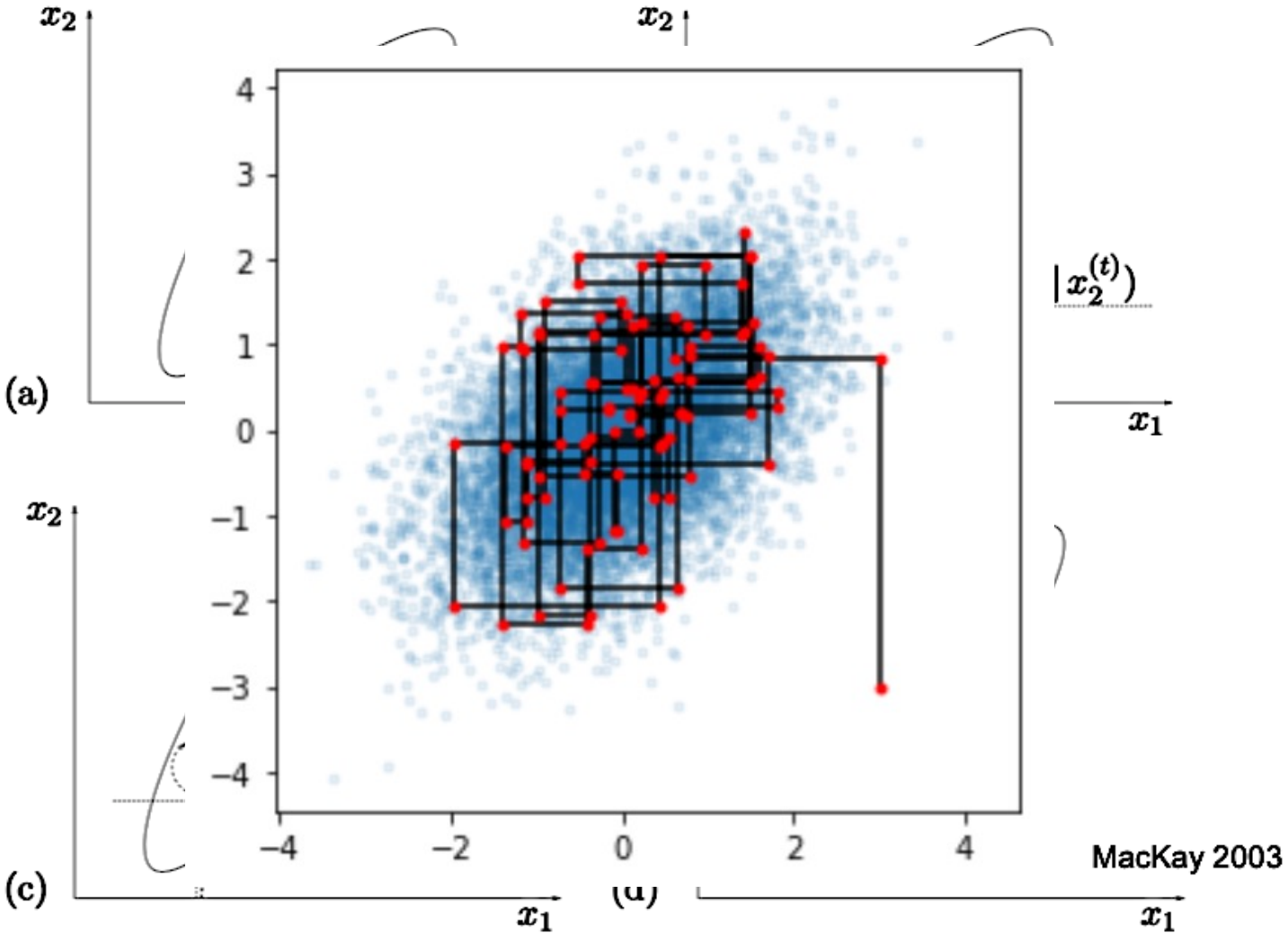
- 0. Select an initial starting point, randomly jump a few steps with a step size;
- 1. Based on θ_n, find θ_{n+1} by evaluating proposal density $q(\theta_{n+1}|\theta_n)$;
- 2. Calculate $a(\theta_{n+1}|\theta_n)$;
- 3. When $a = 1$, accept θ_{n+1}; otherwise, accept θ_{n+1} by a probability a;
- 4. If θ_{n+1} is accepted, $\theta_{n+1} = \theta_{n+1}$; if not, $\theta_{n+1} = \theta_n$;
- 5. repeat the first step.

```python
import numpy as np
import matplotlib.pyplot as plt

# Target distribution: Univariate Gaussian (mean = 5, standard deviation = 2)
def target_distribution(x):
    return np.exp(-0.5 * ((x - 5) / 2)**2) / (2 * np.pi * 2**2)**0.5

# Metropolis-Hastings algorithm
def metropolis_hastings(iterations, proposal_std):
    samples = []
    current_sample = np.random.randn()  # Initial sample from a standard normal distribution

    for _ in range(iterations):
        # Propose a new sample from a Gaussian distribution centered at the current sample
        proposed_sample = current_sample + np.random.normal(scale=proposal_std)

        # Calculate acceptance ratio
        acceptance_ratio = min(1, target_distribution(proposed_sample) / target_distribution(current_sample))

        # Accept or reject the proposed sample based on the acceptance ratio
        if np.random.rand() < acceptance_ratio:
            current_sample = proposed_sample

        samples.append(current_sample)

    return samples

# Parameters
iterations = 10000
proposal_std = 1.0

# Generate samples using Metropolis-Hastings algorithm
samples = metropolis_hastings(iterations, proposal_std)

# Plot the histogram of generated samples and the target distribution
plt.hist(samples, bins=50, density=True, alpha=0.6, label='Metropolis-Hastings Samples')
x_range = np.linspace(min(samples), max(samples), 100)
plt.plot(x_range, target_distribution(x_range), 'r', label='Target Distribution')
plt.legend()
plt.xlabel('Sample Value')
plt.ylabel('Density')
plt.title('Metropolis-Hastings Sampling')
plt.show()
```

GIBBS SAMPLING

- A method for sampling from distributions over at least two dimensions;
- It can be viewed as a special case of the Metropolis method, in which a sequence of proposal distributions Q are defined in terms of the conditional distributions of the joint distribution $P(x)$;
- It assumed that, although $P(x)$ is too complex to draw samples from directly, its conditional distribution $P\left(x_i \middle| \{x_j\}_{j \neq i}\right)$ are tractable to work with.

GIBBS SAMPLING



SLICE SAMPLING

- *It can be applied when the target density $P(x)$ can be evaluated at any point x;*
- *Step-size is less important than Metropolis method;*
- *No requirement that the one-dimensional conditional distributions be easy to sample from, like Gibbs sampling;*
- *Similar to rejection sampling, but no requirement for an upper-bounding function.*

CONVERGENCE CRITERION

- Consider M parallel chains ($j = 1 \cdots M$), each chain contains N points ($i = 1 \cdots N$), then the chain element is y_i^j (a point in parameter space);
- Define the mean of the chain:

$$\bar{y}^j = \frac{1}{N} \sum_{i=1}^{N} y_i^j$$

- Define the mean of all chains:

$$\bar{y}^j = \frac{1}{NM} \sum_{i,j=1}^{N,M} y_i^j$$

CONVERGENCE CRITERION

- *Then the variance between chains:*

$$B_n = \frac{1}{M-1}\sum_{j=1}^{M}\left(\bar{y}^j - \bar{y}\right)^2$$

- *The variance within a chain:*

$$W = \frac{1}{M(N-1)}\sum_{ij}\left(y_i^j - \bar{y}^j\right)^2$$

- *Define the quantity:*

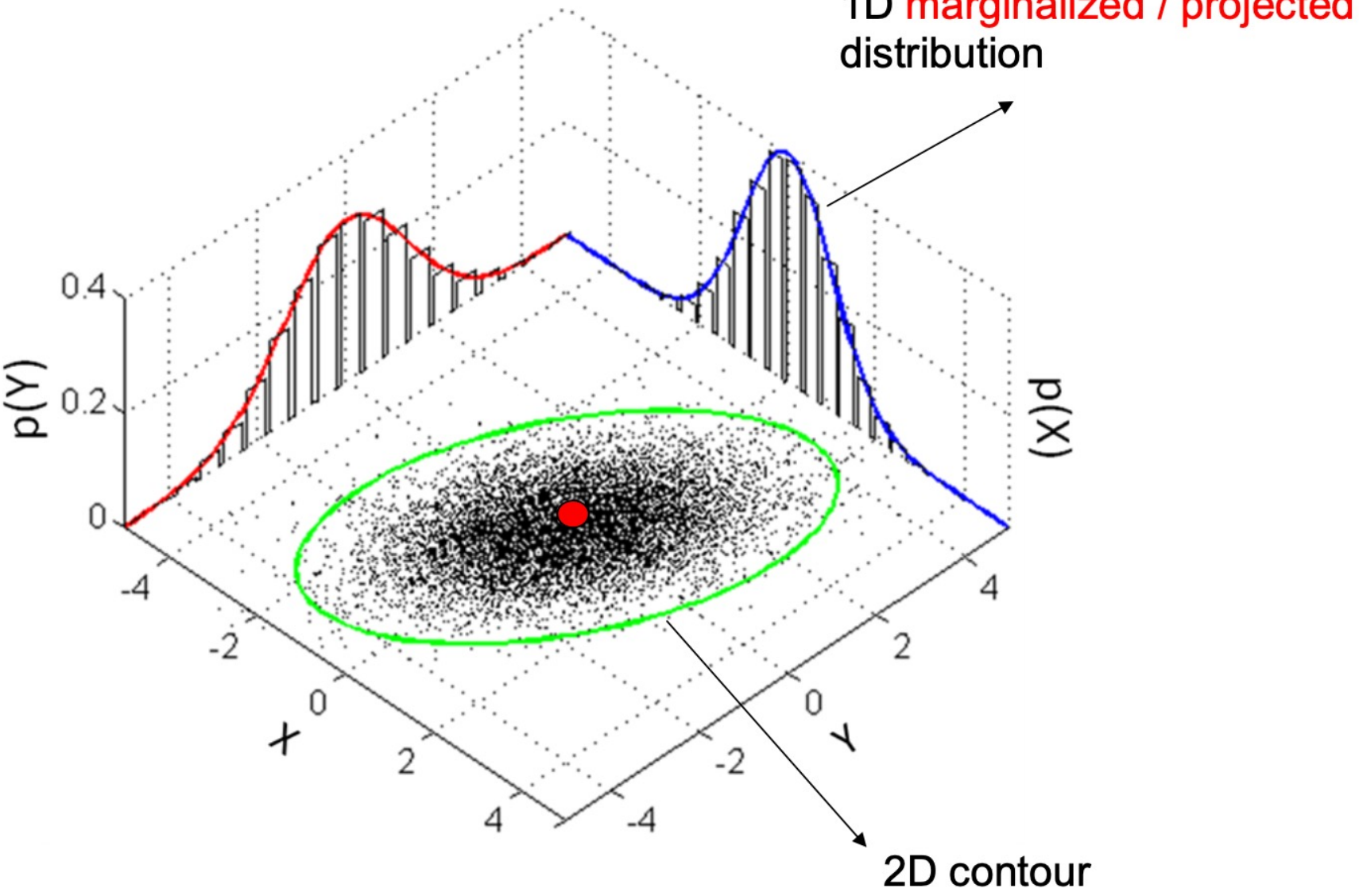
$$\hat{R} = \frac{[N(N-1)]W + B_n(1+1/M)}{W}$$

- *The MCMC chains converge when:*

$$\hat{R} < 1.1$$

AFTER OBTAINING THE MCMC CHAINS

- ***Burn-in period***: *although the Markov chain eventually converges to the desired distribution, the initial samples may follow a very different distribution, especially if the starting point is in a region of low density. As a result, a burn-in period is typically necessary, where an initial number of samples are thrown away (usually the first 100-1000 samples).*
- ***Thinning process***: *after burn-in, in order to obtain independent samples, we should only keep every nth sample in the chains (usually* $n = 10 - 100$*).*
- *Finally, about* $10^4 - 10^5$ *samples or chain points are needed to illustrate the posterior distribution.*


1D marginalized / projected distribution
0.4
0.2
0
p(Y)
-4
-2
0
2
4
X
4
2
0
-2
-4
Y
p(X)
2D contour

THANK YOU