Introduction to Bayesian Methods - 4

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

A short overview of model selection methods

The generic purpose of a model selection statistic is to set up a tension between the predictiveness of a model (for instance indicated by the number of free parameters) and its ability to fit observational data. Oversimplistic models offering a poor fit should of course be thrown out, but so should more complex models that offer poor predictive power.

There are two main types of model selection statistic that have been used in the literature so far. Information criteria look at the best-fitting parameter values and attach a penalty for the number of parameters; they are essentially a technical formulation of "chi-squared per degrees of freedom" arguments. By contrast, the Bayesian evidence applies the same type of likelihood analysis familiar from parameter estimation, but at the level of models rather than parameters. It depends on goodness of fit across the entire model parameter space.

(Liddle & al., 2006 – Astronomy & Geophysics, Volume 47, Issue 4, pp. 4.30-4.33)

Mon. Not. R. Astron. Soc. **377,** L74–L78 (2007) doi:10.1111/j.1745.3933.00306.x

Information criteria for astrophysical model selection

Andrew R. Liddle[1,2★]

[1]*Astronomy Centre, University of Sussex, Brighton BN1 9QH*
[2]*Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, Hawai'i 96822, USA*

Accepted 2007 February 19. Received 2007 February 16; in original form 2007 January 8

ABSTRACT

Model selection is the problem of distinguishing competing models, perhaps featuring different numbers of parameters. The statistics literature contains two distinct sets of tools, those based on information theory such as the Akaike Information Criterion (AIC), and those on Bayesian inference such as the Bayesian evidence and Bayesian Information Criterion (BIC). The Deviance Information Criterion combines ideas from both heritages; it is readily computed from Monte Carlo posterior samples and, unlike the AIC and BIC, allows for parameter degeneracy. I describe the properties of the information criteria, and as an example compute them from *Wilkinson Microwave Anisotropy Probe* 3-yr data for several cosmological models. I find that at present the information theory and Bayesian approaches give significantly different conclusions from that data.

Interlude: the Likelihood Ratio Method and Wilks' theorem – 1

- Taylor expansion close to the true value of the parameter(s)

$$\ln L(D|\theta) \sim \frac{1}{2} \left. \frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right|_{\theta=\theta_0} (\theta - \theta_0)^2 \approx -\frac{1}{2}\mathbb{E}\left[\left| \left. \frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right|_{\theta=\theta_0} \right| \right] (\theta - \theta_0)^2$$

- Integration

$$L(D|\theta) \approx L_{\max} \exp\left\{ -\frac{1}{2}\mathbb{E}\left[\left| \left. \frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right|_{\theta=\theta_0} \right| \right] (\theta - \theta_0)^2 \right\}$$

- Extension to more than one parameter (with parameters split into two subsets)

$$L(D|\boldsymbol{\theta}) = L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \propto \exp\left[-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T I (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right]$$

where Fisher's information matrix is split into submatrices $I = \begin{pmatrix} I_{rr} & I_{rs} \\ I_{sr} & I_{ss} \end{pmatrix}$

Interlude: the Likelihood Ratio Method and Wilks' theorem – 1

- Then, $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_r \\ \boldsymbol{\theta}_s \end{pmatrix}$ and therefore

$$L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})\right.$$
$$\left. -(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rs}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{0,s}) - \frac{1}{2}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{0,s})^T I_{ss}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{0,s})\right]$$

- When we maximize the likelihood with respect to the whole parameter vector, we find that the estimators for the subvectors are

$$\boldsymbol{\theta}'_r = \hat{\boldsymbol{\theta}}_r; \quad \boldsymbol{\theta}'_s = \hat{\boldsymbol{\theta}}_s$$

and the corresponding maximum likelihood has a fixed value that depends only on data.

Interlude: the Likelihood Ratio Method and Wilks' theorem – 1

- When we maximize the likelihood with respect to the s parameters only, we find

$$L(D|\boldsymbol{\theta}_r, \hat{\boldsymbol{\theta}}_s) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r}) - (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rs}(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_{0,s})\right]$$
$$\propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - \mathbf{b}_D)^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - \mathbf{b}_D)\right]$$

- This means that the statistic

$$\lambda = -2 \ln L(D|\boldsymbol{\theta}_r, \hat{\boldsymbol{\theta}}_s)$$
$$\sim (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - \mathbf{b}_D)^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - \mathbf{b}_D)$$
$$\approx (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})$$

(where the bias vanishes asymptotically) has a chi-square distribution with *r* degrees of freedom for large n (Wilks' theorem).

A short overview of model selection methods – ctd.

Akaike Information Criterion (AIC).

This was derived by Hirotugu Akaike in 1974, and takes the form

$$\mathrm{AIC} = -2 \ln \mathcal{L}_{\max} + 2k$$

where k is the number of parameters in the model. The subscript "max" indicates that one should find the parameter values yielding the highest possible likelihood within the model. ***This second term acts as a kind of "Occam factor"; initially, as parameters are added, the fit to data improves rapidly until a reasonable fit is achieved, but further parameters then add little and the penalty term 2k takes over.*** *The generic shape of the AIC as a function of number of parameters is a rapid fall, a minimum, and then a rise. The preferred model sits at the minimum.*

The AIC was derived from information-theoretic considerations, specifically an approximate minimization of the Kullback–Leibler information entropy which measures the distance between two probability distributions.

(Liddle & al., 2006)

Outline of Akaike's derivation

1. max log-likelihood ratio between conjectured model (*k*-dimensional parameter vector) and true model (*l*-dimensional parameter vector)

$$\ln \frac{f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)})}{f(x|\theta)}$$

2. this depends on the dataset, which is distributed according to the true model; to get rid of the fluctuations, we average the max log-likelihood ratio over the true distribution

$$\mathbb{E}\left[\ln \frac{f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)})}{f(x|\theta)}\right] = \int_{\Theta} f(x|\theta) \ln \frac{f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)})}{f(x|\theta)} dx = -D_{\mathrm{KL}}\left(f(x|\theta)||f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)})\right)$$

3. here we remark that:
   - this is purely theoretical, since we do not know the true pdf
   - the r.h.s. expression is the negative of the Kullback-Leibler divergence between the conjectured and the true pdf
   - the l.h.s. is maximum when the KL divergence is at a minimum
   - the r.h.s. expression can be written as

$$\int_{\Theta} f(x|\theta) \ln \frac{f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)})}{f(x|\theta)} dx = \int_{\Theta} f(x|\theta) \ln f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)}) dx - \int_{\Theta} f(x|\theta) \ln f(x|\theta) dx$$

Outline of Akaike's derivation

4. the second term in the expansion is unknown, but it is a constant and we can get rid of it, and change sign as well (with an additional factor 2, see later), so that by minimizing the first term we actually minimize the KL divergence

$$\int_\Theta f(x|\theta) \ln \frac{f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)})}{f(x|\theta)} dx = \int_\Theta f(x|\theta) \ln f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)}) dx - \int_\Theta f(x|\theta) \ln f(x|\theta) dx$$

$$\rightarrow -2 \int_\Theta f(x|\theta) \ln f(x|\hat{\theta}^{(k)}, \theta^{(|\ell-k|)}) dx$$

5. going back to Wilks' theorem, we know that the remaining |*l*-*k*| degrees of freedom in the likelihood ratio are (asymptotically) normally distributed, therefore the -2log has a chi-square distribution with |*l*-*k*| degrees of freedom, with mean value |*l*-*k*|, and therefore the required mean value has an asymptotic bias 2|*l*-*k*|;
6. using the max likelihood as an estimator of the mean, we find that the discrepancy expressed by the equation above can be written as

$$-2 \ln f(x|\hat{\theta}^{(k)}) + 2k$$

after dropping the constant l

Bayesian Information Criterion (BIC).

This was derived by Gideon Schwarz in 1978 and strongly resembles the AIC. It is given by

$$\mathrm{BIC} = -2 \ln \mathcal{L}_{\max} + k \ln N$$

where N is the number of datapoints. Since a typical dataset will have $\ln N > 2$, the BIC imposes a stricter penalty against extra parameters than the AIC.

It was derived as an approximation to the Bayesian evidence**, to be discussed next, but **the assumptions required are very restrictive and unlikely to hold in practice, rendering the approximation quite crude.

(Liddle & al., 2006)

Bayesian evidence

Model selection aims to determine which theoretical models are most plausible given some data, without necessarily considering preferred values of model parameters.

Ideally, we would like to estimate posterior probabilities on the set of all competing models using Bayes' theorem:

$$P(M_i|D,I) = \frac{P(D|M_i,I)P(M_i|I)}{\sum_k P(D|M_k,I)P(M_k|I)}$$

and select the best model using the odds ratio

$$\mathcal{O}_{i,j} = \frac{P(M_i|D,I)}{P(M_j|D,I)} = \frac{P(D|M_i,I)P(M_i|I)}{P(D|M_j,I)P(M_j|I)}$$

or the Bayes factor, if we assume equal prior probabilities for the different models:

$$B_{i,j} = \frac{P(D|M_i,I)}{P(D|M_j,I)}$$

Thus, we see that the Bayes factor is a ratio of evidences

$$B_{i,j} = \frac{P(D|M_i, I)}{P(D|M_j, I)}$$

As usual, each evidence is obtained by marginalizing the likelihood with respect to the (potentially different) parameters:

$$P(D|M_i, I) = \int_{\Theta_i} P(D|\boldsymbol{\theta}_i, M_i, I) p(\boldsymbol{\theta}_i|M_i, I) d\boldsymbol{\theta}_i$$

The evidence of a model is … the average likelihood of the model in the prior.

Unlike the AIC and BIC, it does not focus on the best-fitting parameters of the model but asks "of all the parameter values you thought were viable before the data came along, how well on average did they fit the data?". Literally, it is the likelihood of the model given the data.

The evidence rewards predictability of models, provided they give a good fit to the data, and hence gives an axiomatic realization of Occam's razor.

A model with little parameter freedom is likely to fit data over much of its parameter space, whereas a model that could match pretty much any data that might have cropped up will give a better fit to the actual data but only in a small region of its larger parameter space, pulling the average likelihood down.

(Liddle & al., 2006)

Which statistics?

Of these statistics, we would advocate using – wherever possible – the Bayesian evidence, which is a full implementation of Bayesian inference and can be directly interpreted in terms of model probabilities. It is computationally challenging to compute, being a highly peaked multidimensional integral, but recent algorithm development has made it feasible in cosmological contexts.

If the Bayesian evidence cannot be computed, the BIC can be deployed as a substitute. *It is much simpler to compute as one need only find the point of maximum likelihood for each model.* ***However, interpreting it can be difficult. Its main usefulness is as an approximation to the evidence, but this holds only for gaussian likelihoods and provided the datapoints are independent and identically distributed.*** *The latter condition holds poorly for the current global cosmological dataset, though it can potentially be improved by binning of the data, hence decreasing the N in the penalty term.*

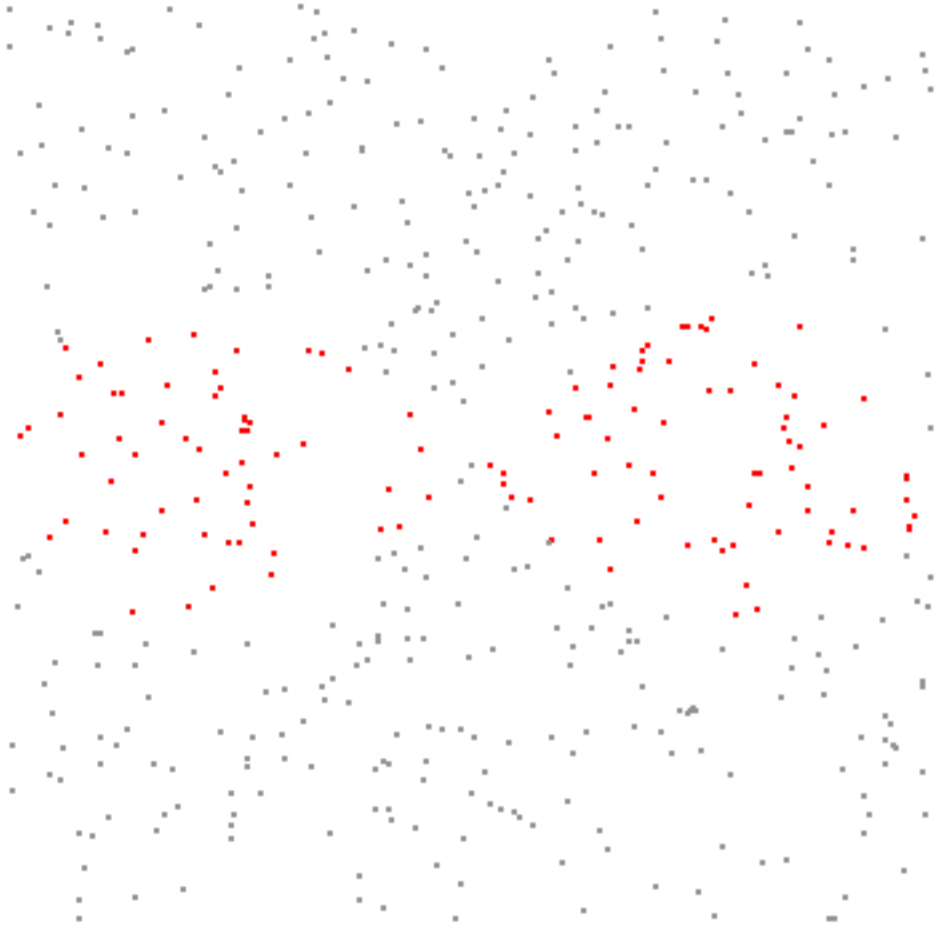
The AIC has been widely used outside astrophysics but is of debatable utility. ***It has been shown to be "dimensionally inconsistent", meaning that it is not guaranteed to give the right result even in the limit of infinite unbiased data.*** *It may be useful for checking the robustness of conclusions drawn using the BIC.* ***The evidence and BIC are dimensionally consistent.***

(Liddle & al., 2006)

Back to basics: what if we “measure” a mathematical constant instead of a physical parameter?

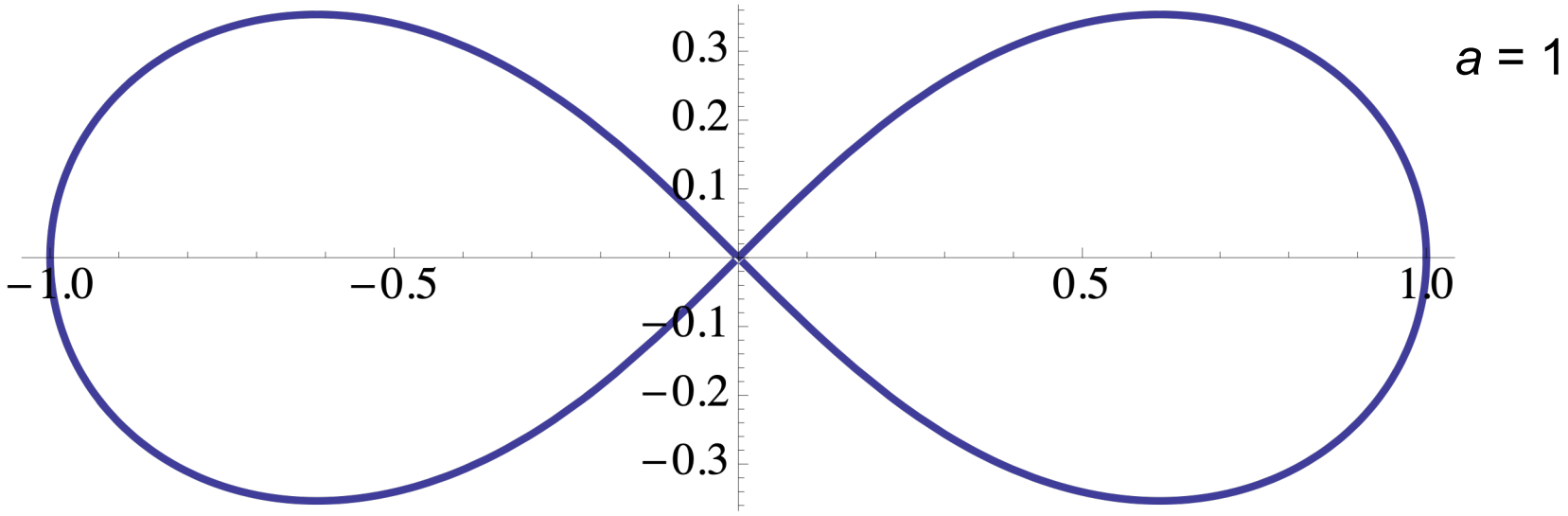
***Example**:*

area of Bernoulli’s lemniscate obtained with a Monte Carlo simulation.



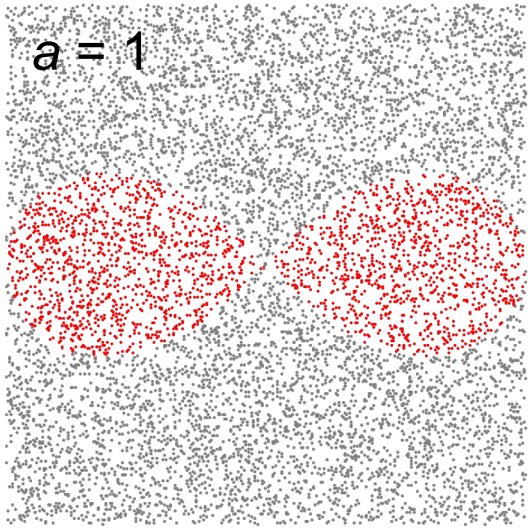
Parametric equation of Bernoulli's lemniscate

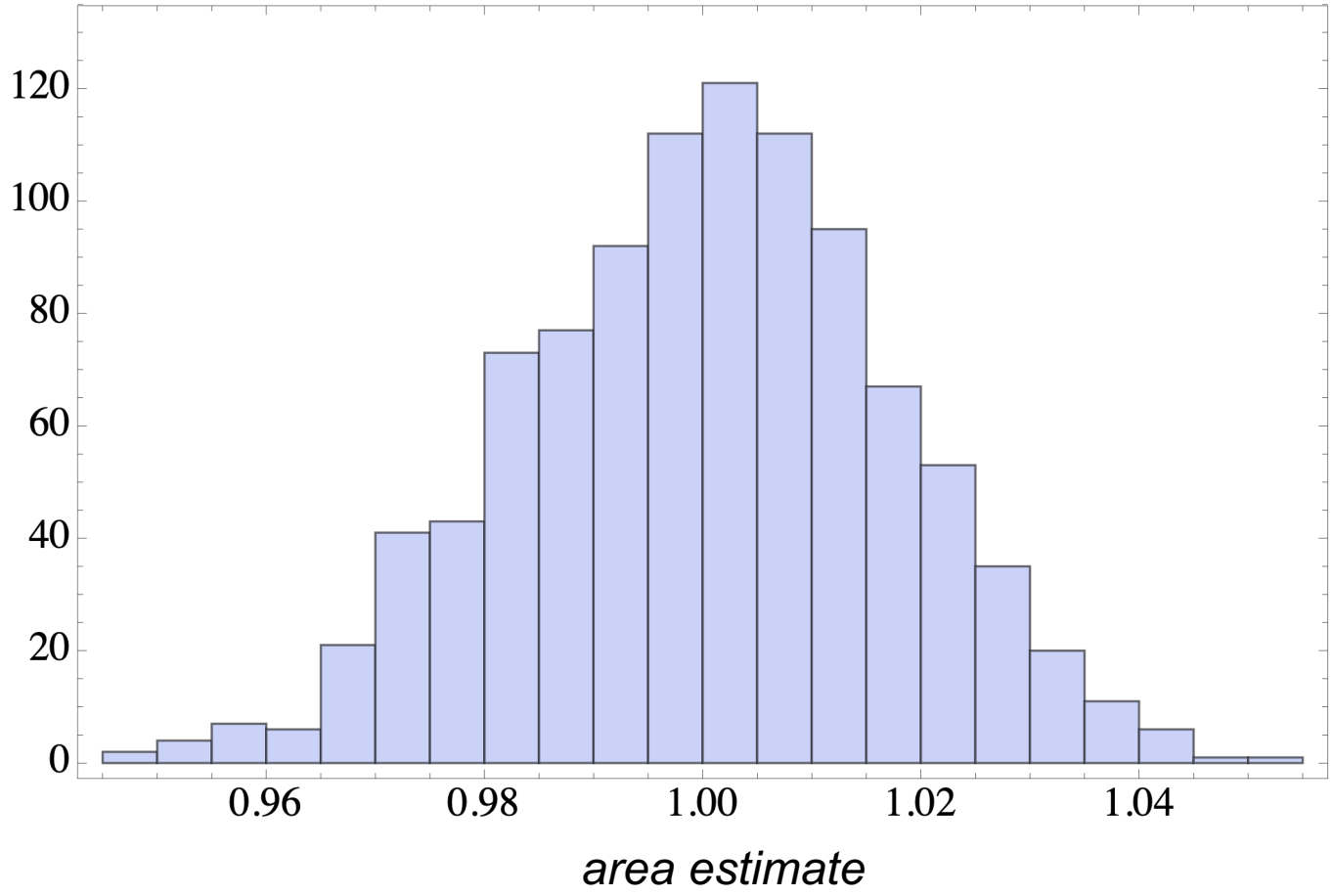
$$r = a\sqrt{\cos 2\theta}$$



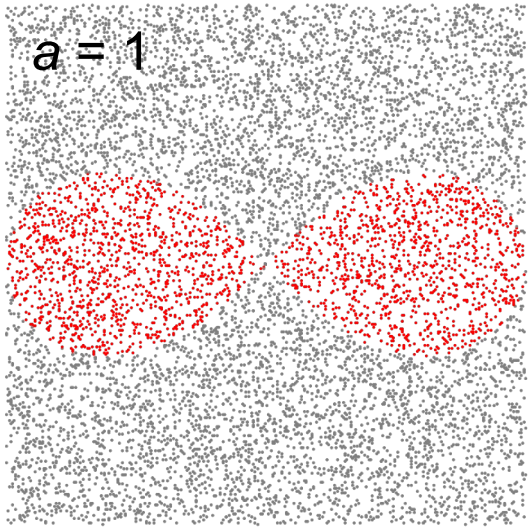
What is its area?

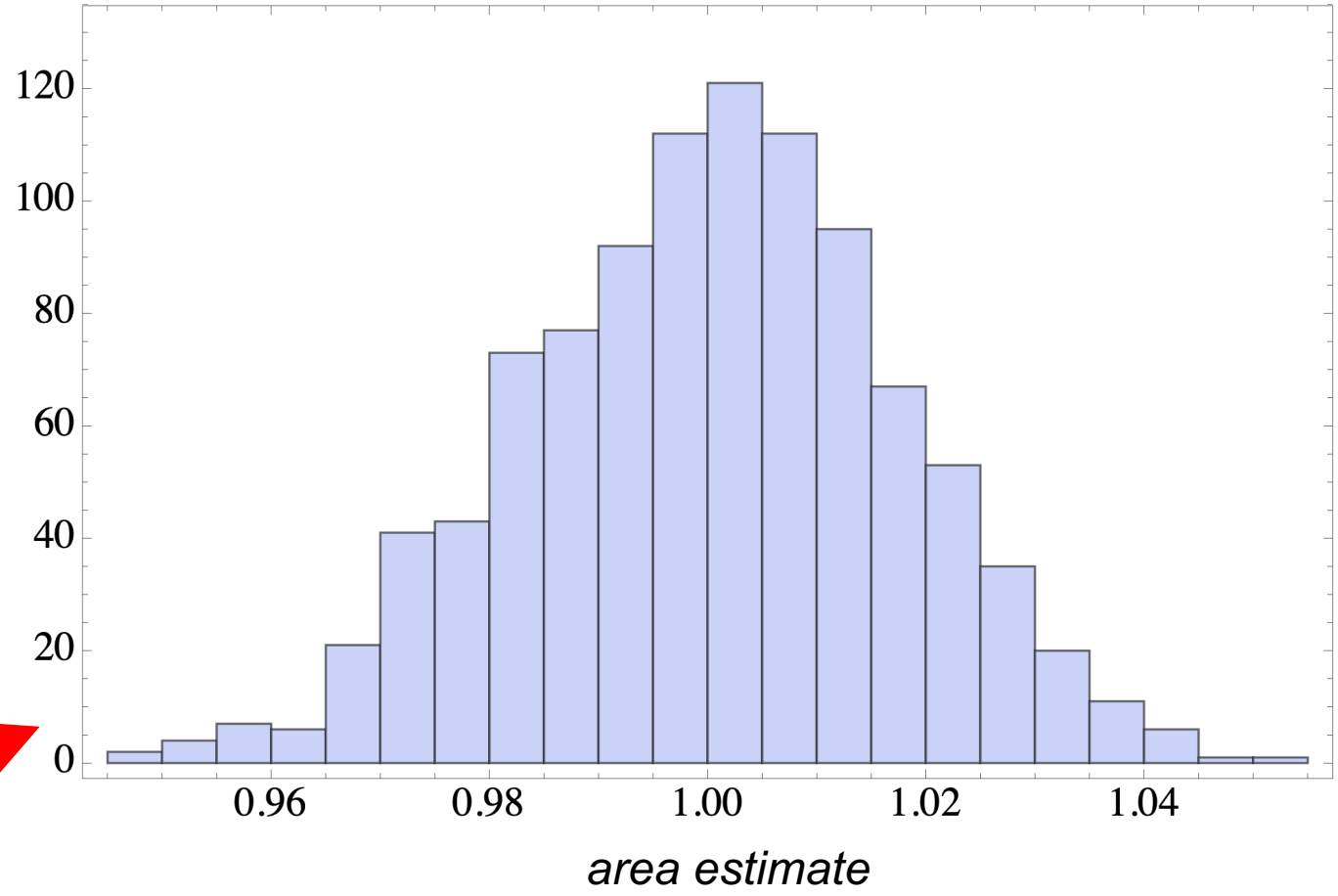
Empirical Monte Carlo distribution of the area estimate





Empirical Monte Carlo distribution of the area estimate

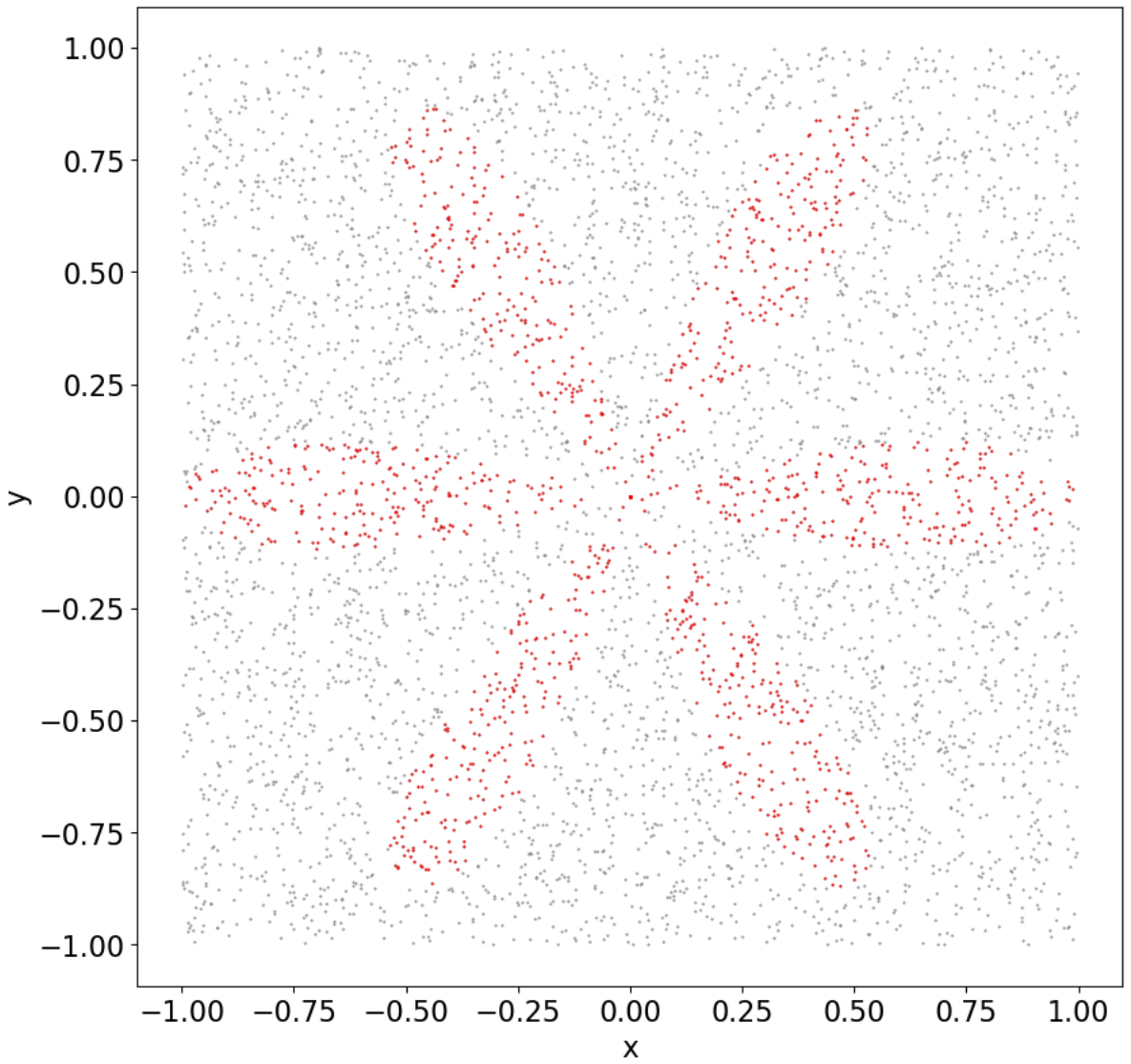




a probability distribution of
a mathematical constant???

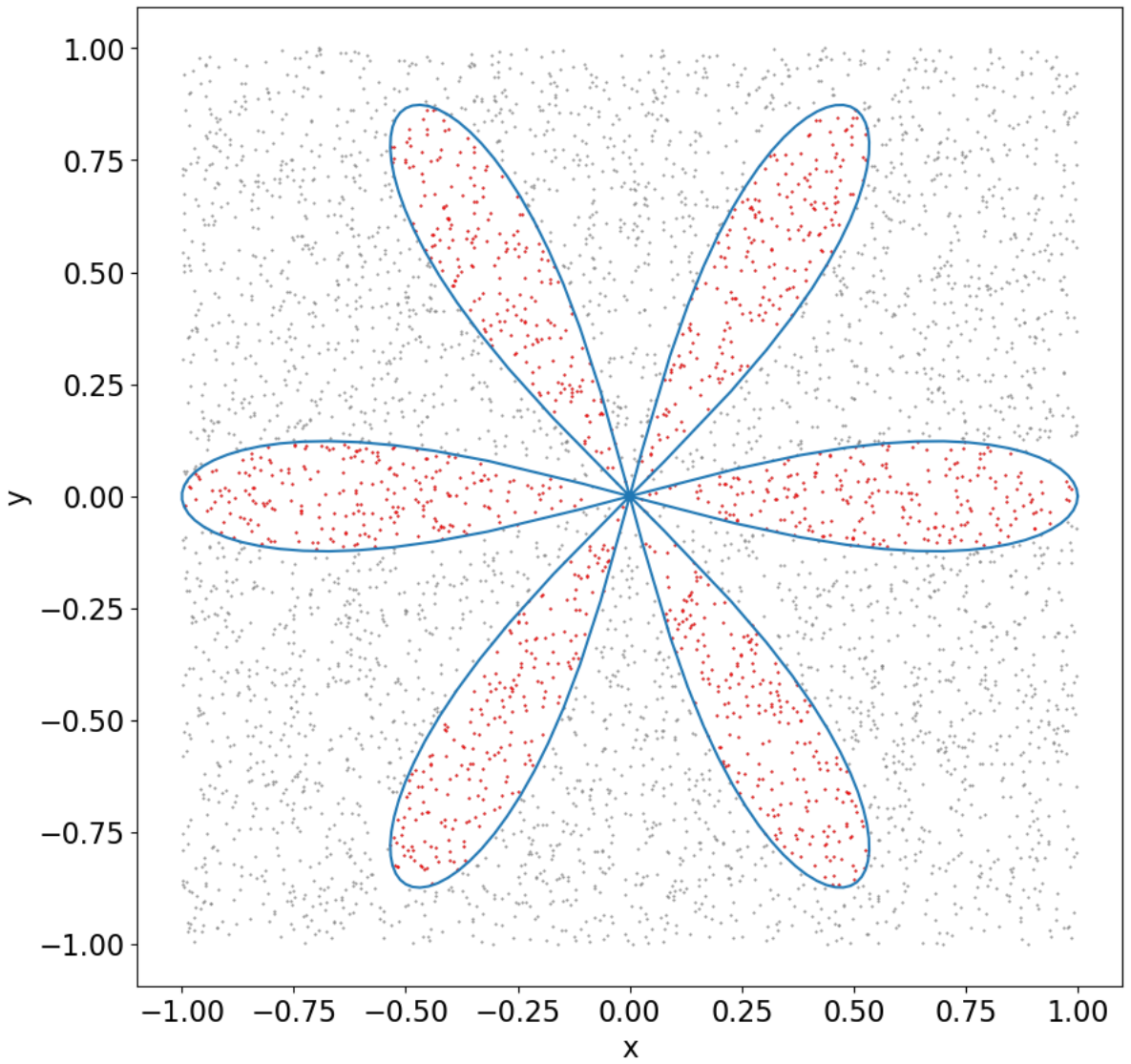
Now, try yourself with a rhodonea

$$r^2(\theta) = a^2 \cos(6\theta)$$



Now, try yourself with a rhodonea

$$r^2(\theta) = a^2 \cos(6\theta)$$



AMERICAN
JOURNAL of PHYSICS

A Journal Devoted to the Instructional and Cultural Aspects of Physical Science

VOLUME 14, NUMBER 1 JANUARY–FEBRUARY, 1946

Probability, Frequency and Reasonable Expectation

R. T. Cox
The Johns Hopkins University, Baltimore 18, Maryland

See also R T Cox, *The Algebra of Probable Inference*, The John Hopkins Press (Baltimore, 1961)
https://bayes.wustl.edu/Manual/cox-algebra.pdf

Boolean algebra (symbolic logic)

a, **b**, **c** ... propositions (true or false)

Basic operations

OR: **a** ∨ **b**

AND: **a**·**b**

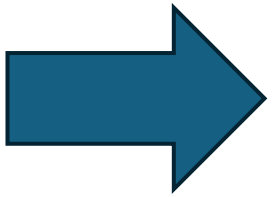
NOT: **~a**

Truth tables

a	b	a ∨ b
T	T	T
T	F	T
F	T	T
F	F	F

a	b	a · b
T	T	T
T	F	F
F	T	F
F	F	F

a	~a
T	F
F	T



see handout