Lect. 17 → Renamed Lect 17 (2023dec03) | 2021nov22 | Updated 2023dec2 | 1 | 1700 1

Bayesian Analysis (BA)

(A name I prefer to the usual Bayesian Inferenc

1) Preamble: BA is path to truth quantified a Scientific method Quantified. (P. 2

2) Unquantified or Qualitative Bayesian Analysis Priors & Posteriors (P. 3)

3) Bayes Theorem & a bit of History (P. 6

4) Proof of Bayesian Analysis is in the ideal limit — which can be approached arbitrarily closely (P. 11

5) Bayes Odds Ratio & Bayes Factor (P. 21

6) ~~Truth~~

7) Multinomial Theorem & Multinomial Probability (P. 23)

8) An Example of Bayesian Analysis
A toy example: The Dice Problem

9) Marginalization & Occam's Razor
shortcuts to Bayesian analyses

10) BIC = Bayesian Information Criteria
AIC = Akaike Information Criterion

~~2~~ 7002

7.1

Preamble

Note in this lecture I give a proof Bayesian analysis. Not an explication of how to do it in practice — immense amount of tricks, procedures, formulas → but there are computer packages

I argue ~~that~~ Bayesian Analysis (BA = AKA Bayesian inference or Bayesian statistics) is a true theory of nature truth or i.e., a true theory in an ideal limit that can be approached arbitrarily closely ideally.

many important theories are like their E.g., Newtonian Physics, ↪ It is believed to be the macroscopic, low velocity limit of QM ~~limit of~~ Special relativity weak gravity limit of General relativity

at least improvement, The Scientific Method Quantifies and thereby the proof of the Sci Method

exact virtually all, except maybe Theory of everything & 2nd law of Thermodynamics) ~~are~~

Another perspective is that Newtonian physics is an approximate theory

absolutely exact. But maybe their exactness is quibbling

But yours truly (following a head) would call it a true emergent theory → exactly true in the classical limit which can be approached arbitrarily closely and is approached very closely

in everyday life.

Other examples 2nd law of thermodynamics (a law in qualitative sense true by reco

Maybe actually not? Well the system must be big enough

Natural selection evolution

↳ unlike Newtonian Physics ⇒ these can be proven by mathematical logic (and law? Well for my toy univ true)

— and so can Bayesian Analysis

↳ aside from pure rigor or pure philosophical quibbling in my opinion.

17.2) Qualitative Bayesian Analysis

⇒ We do it all the time and so all other conscious beings to one degree or another

All of life experience gives vague probabilities about what [things] will happen in a given situation

e.g., impossible, virtually impossible, highly unlikely, unlikely, possibly, likely, very likely, virtually certain, certain

17004

Based on vague approximate frequency

N_{event}

N_{trial}

on the set of stakes

initiative

These are your prior probabilities on priors

Then New experience happens in that situation and you vaguely update your priors to your posteriors (which sounds awful and obtuse) {i.e. posterior probabilities

your updateing is also qualitative and of variable accuracy, but it works well enough mostly in everyday life to your advantage

Nature via Natural selection evolution has its own way improving success rate relative to local conditions.

Of course, things like major impactors can change the rules of success suddenly → like what happened to non-avian dinosaurs

17005

Ex. Job interviews

↳ especially if you are new at the game — or the 'rules' have changed ⇒ each interview causes you to update your priors to posteriors often in an intuitive (but still useful) way.

To some degree 'qualitative Bayesian' analysis trial & error improvement if just restricted to each new experience ~~of a unique situation~~ ~~from~~ has very restricted generality.

But as more general conclusions about updates occur ⇒ qualitative Bayesian analysis approaches the scientific method.

2006

Bayesian analysis itself is the sci. Method Quantified (I argue and the proof therefore of the scientific method)

7.3) Bayes Theorem {At root of BA and general to all probability theory in fact}

Root of Bayesian analysis is Bayes theorem — which is really simple and simple to prove. Easier to prove than to remember.

Discovered by Thomas Bayes (c.1701–1761) and published posthumously in 1763. (Wik)

Pierre-Simon Laplace (1749–1827) independently discovered it (and published? it) in 1774 (Wik)

~~Consider 3 events A, B, K~~

~~N is the background knowledge that I introduce as extra item because I need it later. It isn't needed for the proof of Bayesian theorem itself.~~

~~I use ABK as short-hand for "and" — we say shorthand for union = U symbol in math which is too klutzy for me in this context~~

order has no meaning.

2025 May 01

~~7~~ 17007

Conditional Probability and Probability Product Rule

Consider Events A, B, K <Need K for proof of Bayesian Analysis (BA)

ABK is ~~overlap~~ joint event intersection all 3 at once

$A \cap B$ intersection of both A and B

$A \cup B$ union of sets A and B

$ABK = A \cap B \cap K$ so NOT a product

$\cap$ symbol, intersection

Background knowledge that in BA jargon is often called conditioning knowledge or context, (Loredo 2024 p. 3) but too klutzy for a shorthand and looks ugly in general

$$P(ABK) = \frac{N_{ABK}}{N} = \frac{N_{ABK}}{N_{BK}} \frac{N_{BK}}{N_K} \frac{N_K}{N}$$

Probability of all 3 at once - order has no meaning

Frequentist definition N_{ABK} cases of N trials taken to infinity

Nothing forbid this factorization

$$P(ABK) = P(A|BK)P(B|K)P(K)$$

BK is also given B K implicitly

By symmetry

$$P(ABK) = P(B|AK)P(A|K)P(K)$$

∴ ~~Bayes theorem in symmetrical~~ form

$$P(A|BK)P(B|K) = P(B|AK)P(A|K) \quad \text{QED}$$

easier to prove than to remember.

Asymmetric form

$$P(A|BK) = \frac{P(B|AK)P(A|K)}{P(B|K)}, \quad P(B|AK) = \frac{P(A|BK)P(B|K)}{P(A|K)}$$

Usually show with K implicit, but for proof of BA, we need K explicit.

Some assert this is an axiom or can be so taken. (Twitter ? p ?)

But I believe Probability only has meaning in a Frequentist sense even if N_{ABK}/N is only vaguely known or hypothesized imaginatively. In fact, BA means you can improve from vaguely known to "truth"

N_{ABK} & N exist in some ideal limit as Galileo would argue.

8
7008

That's all: a profound and profoundly simple logical rule of reality

Root of Bayesian analysis

Bayesian analyses was greatly elaborated in 1939 on (Scientific Inference 3rd ed 1973) by Harold Jeffreys (1891–1989) (no relation — but he manages to look like my father anyways)

What is Bayesian analysis?

It's a path to true theories by using Bayes theorem to update prior probabilities for posterior probabilities for theories until one true theory is left.

In toy cases this is easy. In hard (or real) cases, hard, because there is a lot of slop in assigning priors, [Not to theories in almost cases but the priors on theory parameters.]

A theory with free parameters in one sense is a continuum of theories

No one much used Bayesian analysis before 1990s. Since then it has had growing vogue.

...lly, ...dual ...tive. ...ideal ...ps ...cond ...p 1 ... not ...ne. ...eryone ...ems to ...nt step 1 ... ...th the ...e that ...ep 1's ...ll ...actually ...et better ...y

...one in single ...pers. 'Now we ...d this data, now ...s and take ...way that'

17009

Why was it unused before?

Well, its real power is in dealing with statistical theories (i.e., theories that only give probabilities) and interesting cases in cosmology, epidemiology, social sciences only came with vast data sets and vast computing power to manipulate them. But since 1990s ~~[redacted]~~ on, we have enough of both.

So nowadays, Bayesian analysis is an ubiquitous tool (though less wonderful than one can hope).

You may ask how can a theory have a probability of being true, isn't it just true or false

~~to~~
17010)

speaking in an absolute sense not leaving aside partially
ive theories ~~or~~ having some truth and some
falseness definitely in
less absolute sense not leaving aside partially true theories

But the probability of truth
is NOT in the absolute
sense.

It's probability to our
knowledge

Example

I've a coin in my hand.
Is it heads or tails?

It's one or the other
in absolute sense.

But to your knowledge
it's 50% - 50%

See. Nothing in my hand
→ So you should include
the theory of deception.

17.4) Proof that Bayesian Analysis in the Ideal Limit is the Path to True Theories But You May Need to Restrict their Realm of Validity

a) In yours truly's view, Bayesian analysis (BA) is the scientific method quantified.

Many probably share this view, but many probably disagree since philosophers of science are great quibblers.

The proof is NOT absolutely rigorous NOR beyond quibbling.

Say you have background knowledge which in jargon of Loredo & Wolpert (2024, p. 3) is called context C_0 or conditioning information which is all knowledge relevant to the aspect of reality whose true theory is the goal of your BA. C_0 includes general theories, specific information, anything at all relevant.

From C_0, you deduce a set of theories $\{T_i\}$ which could be considered already in C_0 or adding to C_0. The set $\{T_i\}$ does NOT have to be complete even to C_0 as it is, but in practice

7012)

it should be all theories you think most probable by qualitative Bayesian analysis

Completeness is a rather vague concept any way. To recapitulate, the power of BA is that it works in the ideal limit no matter how much vagueness there is in choosing $\xi T_i \xi$, assigning their initial probabilities, the multiplicative factor frequentist probability assumption, etc.

Note, the theories are assumed to be a finite discrete set. They have no free parameters. A theory with free parameters is in a sense a continuum family of theories. Of course, many theories have free parameters. For these, you need marginalization. see p. 170

Note, $\xi T_i \xi$ does NOT have to include the true theory but if it does NOT, you will have do one or more episodes of qualitative Bayesian analysis to determine a new set of most probable theories.

b) The procedure of ideal BA analysis requires a series of new data acquisitions;

$$D_1, D_2, D_3 \ldots D_e \ldots$$

which I think does NOT have to be infinite

provided all knowledge relevant to the analyzed aspect of reality somehow.

Note D_ℓ can be measurements, but also newly acquired knowledge or theories about some other aspects of reality

The 0th step in the procedure is "initial probabilities" to the set of theories $\{T_i\}$

initial probabilities = initial priors

$$P(T_i \mid C_0)$$

conditional probability of T_i given context C_0

T_i: theory T_i

C_0: context

How do you do this?

In principle, anyway you like.

The ideal BA works no matter how.

But since you thought $\{T_i\}$ was the most probable set and may NOT have any reasonable quantitative ranking, you can use the Principle of indifference (Wik).

Just set all $P(T_i \mid C_0)$ equal.

You can normalize the probabilities if you like!

$$\sum_i P(T_i \mid C_0) = 1$$

but this isn't necessary since ideal BA normalizes updated theory probabilities automatically, (see p. 1701)

As you acquire the new sets of data $D_1, D_2, D_3, \ldots D_\ell, \ldots,$ the context updates

$$C_\ell = C_{\ell-1} D_\ell$$

Where as a shorthand I use 'product' to mean union: i.e.

$$C_{\ell-1} D_\ell = C_{\ell-1} \cup D_\ell$$

Union of sets (wiki)

I think this symbol klutz X typed or handwritten.

You use D_ℓ to update your prior probabilities at step ℓ to update your prior probabilities at step ℓ (i.e. $P(T_i | C_{\ell-1})$) to your posterior probabilities at step ℓ (i.e. $P(T_i | C_\ell)$)

which are your priors for step $\ell+1$.

c) Updating Priors to Posteriors using Bayes theorem at step ℓ

Consider joint probability and factorize

$$P(T_i\, C_{\ell-1}\, D_\ell)$$

$$= P(T_i \mid C_{\ell-1} D_\ell) P(C_{\ell-1} D_\ell) = P(T_i \mid C_\ell) P(D_\ell \mid C_{\ell-1}) P(C_{\ell-1})$$

$$= P(D_\ell \mid T_i C_{\ell-1}) P(T_i \mid C_{\ell-1}) P(C_{\ell-1})$$

$$\therefore \quad P(T_i \mid C_\ell) P(D_\ell \mid C_{\ell-1}) = P(D_\ell \mid T_i C_{\ell-1}) P(T_i \mid C_{\ell-1})$$

$P(C_{\ell-1})$ cancel out, but they are also 1 since we have (i.e. know) $C_{\ell-1}$

which is just the symmetric Bayes theorem from p. 17017 recovered

$$P(T_i \mid C_\ell) = \frac{P(D_\ell \mid T_i C_{\ell-1}) P(T_i \mid C_{\ell-1})}{P(D_\ell \mid C_{\ell-1})}$$

Posterior for step ℓ

Prior for step ℓ

$P(D_\ell \mid T_i C_{\ell-1})$ is the marginal likelihood of the data D_ℓ

17016

Actually, $P(D_\ell \mid T_i\, C_{\ell-1})$ is NOT from a marginalization for our case of a discrete set of theories with no free parameters, but for consistency when do marginalization to get $P(D_\ell \mid T_i C_{\ell-1})$ (see p. 17 ...) we call it that here following the jargon of Kass & Raftery (1995, p. 776).

But what is denominator $P(D_\ell \mid C_{\ell-1})$? We can only do an estimate based on the set of theories $\{T_i\}$ we are considering. We expand using factorization

$$P(D_\ell \mid C_{\ell-1}) = \sum_j \frac{N_{D_\ell}}{N_{T_j C_{\ell-1}}} \frac{N_{T_j C_{\ell-1}}}{N_{C_\ell}}$$

$$= \sum_j P(D_\ell \mid T_j\, C_{\ell-1})\, P(T_j \mid C_{\ell-1})\, P(C_{\ell-1})$$

$$= \sum_j P(D_\ell \mid T_j\, C_{\ell-1})\, P(T_j \mid C_{\ell-1})$$

really just a best estimate to your knowledge. You do NOT know $P(D_\ell \mid C_{\ell-1})$ in an absolute sense.

But this is 1 since we know it.

You could include theories NOT in your set $\{T_i\}$, but that seems pointless practically AND as a formalism

$$\therefore P(T_i|C_e) = \frac{P(D_e|T_i C_{e-1}) P(T_i|C_{e-1})}{\sum_j P(D_e|T_j C_{e-1}) P(T_j|C_{e-1})}$$

$$= \left(\frac{P(D_e|T_i C_{e-1})}{\sum_j P(D_e|T_j C_{e-1}) P(T_j|C_{e-1})} \right) P(T_i|C_{e-1})$$

$$= \left[\frac{P(D_e|T_i C_{e-1})}{\langle P(D_e|T_j C_{e-1}) \rangle} \right] P(T_i|C_{e-1})$$

You can regard the coefficient as an updating factor from prior $P(T_i|C_{e-1})$ to posterior $P(T_i|C_e)$

Note the $P(T_i|C_e)$ are automatically normalized as noted on p. 17019: $\sum_i P(T_i|C_e) = 1$.

Weighted mean of the marginal likelihood

If $P(D_e|T_i C_{e-1})$ exceeds/subceeds the mean, the probability of theory T_i increases/decreases,

Except for formal niceness (i.e., knowing it exists and can be calculated), you do NOT need the denominator. You just update the relative probabilities.

Recall the denominator formula is actually just a best estimate to your knowledge

d) As the BA Procedure Continues What Happens

See p. 17016

As data sets D_e continue, the probabilities of the theories will continue be updated.

7018

i) Quasi-ultimately in many cases,
one T_{i*} will grow in probability
heading toward $P(T_{i*}|C_\ell) = 1$
and the other theories
will head toward $P(T_i|C_\ell) = 0$,
some may have reach $P(T_i|C_\ell)=0$

ii) However, unlucky sets of data D_ℓ may cause $P(T_{i*}|C_\ell)$ to decrease sometimes for some steps. (during the procedure and can be dropped then, of course.)

iii) Does T_{i*} reach $P(T_{i*}|C_\ell) = 1$ in finite or infinite steps.
It can reach it in finite steps,
but if the theories $\{T_i\}$
are statistical and you
gather statistical data D_ℓ
$P(T_{i*}|C_\ell)$ may go 1
only as $\ell \to \infty$.

2026 jan 24

(1701)

However, even for statistical theories, you may be able to access data with is not determined probabilistically and reach $P(T_{i*}|C_\ell) = 1$ for finite ℓ.

I give a toy example of the ideal BA procedure on p. 17 in which this happens.

iv) What if at step ℓ all $P(T_i|C_\ell) = 0$?

Then you have do qualitative Bayesian analysis and invent a new set of theories $\{T_i\}$.

Of course, you don't have wait for this catastrophe.

You can drop theories from the procedure when their $P(T_i|C_\ell) << 1$ and not wait for $P(T_i|C_\ell) = 0$.

And the growing C_ℓ may lead you to new theories that seem probable along the way and you can add them with reasonable reassignment of probabilities.

17020

v) Does $P(T_{i*} | C_\ell) = 1$ mean T_{i*} is the true theory?

No. It just means all the other theories have zero probability.

If you keep doing the procedure with just one theory at some data set D_ℓ, $P(T_{i*} | C_\ell) = 0$!

Then you have to do qualititative Bayesian analysis and invent a new set of theories an start over.

ii) But what if C_ℓ is enormous and the $P(T_{i*} | C_{\ell+1}) = 0$,

In this case, you may well want to retain T_{i*}, but just limit its scope to C_ℓ.

2024jan24 17021

In one sense, this is lowering the bar.

But theories T_i* that are adequate for a vast realm C_i usually shouldn't be dismissed.

Example Newtonian physics was once hypothesized to be absolutely fundamental for all phenomena.

But even early people couldn't explain electromagnetism or chemistry by it though they may hoped that that would happen someday.

However, by later 19th century that hope was diminishing.

For example, classical electromagnetism which seemed an extremely robust theory was based on Maxwell's equations which were known NOT to be invariant under the Galilean transformation

Famously, Einstein was led to special relativity by, among other things, the belief that

7022

Maxwell's equations had to be more fundamental than Newtonian physics.

The advent of classical electromagnetism, special relativity, general relativity, quatum mechanics, and quantum field theory vastly limited the scope of Newtonian physics.

However, saying it's just an approximate theory seems inadequate considering its vast realm of applicability.

Follow a hevd, I'd say it is an exact theory in the classical limit where relative velocities $\ll c$, size scale >> microscopic, the local gravity field is sufficiently uniform, and one avoids some tricky electromagnetic effects.

It's a tune or fundamental emergent theory in the classical limit.

Newtonian can never be proven false, it can only maybe become limited.

This is true of other profound theories too.

We believe general relativity fails in the quantum gravity realm even though we don't have an established quantum gravity theory.

2024jan24 | 17023

vii) Does BA guarantee that you will get to a true theory even if only limited to vast C_2?

I think only if you can keep interrogating reality until you know everything relevant to the aspect of reality of interest.

To give an example of case where practically you may never know the true theory.

Say you just measure series 64-bit floating point number (15 to 17 digits) in the range (0, 1) (and so excluding endpoints 0 and 1).

You measure a long range and all ordinary statistical tests find them to be random, distributed over range (0, 1).

But are the numbers fundamentally random (as set by atmospheric noise; i.e., atmospheric radio noise: Wik) or computer generated random by a deterministic algorithm:

'7024

e.g., the Mersenne-Twister (MT)
which has repeat cycle $= 2^{19937} - 1$
(Wik: Mersenne-Twister: characteristics)

Say you had an exaflop computer and it takes 100 flops (just guessing) to compute one MT random number, how long until you complete a cycle?

$$t = \frac{(100 \text{ flops}) * (2^{19937} - 1)}{10^{18} \text{ flops/s}}$$

$$\triangleq \frac{10^2 * 10^{[0.3*(20000)]} \text{ flops}}{10^{18} \text{ flops/s}}$$

$$= \frac{10^{6000}}{10^{18}} \text{ s} \triangleq 10^{6000} \text{ s} \times \left(\frac{1 \text{ year}}{\pi \times 10^7 \text{ s}}\right)$$

$$\triangleq 10^{6000} \text{ years} \triangleq 10^{6000} \text{ Gyr}$$

Practically, you could never tell as long as MT numbers were as good as claimed.

But if you could see the MT numbers were just coming from an isolated computer, you'd know they were algorithm generated.

This example is a pathological case,
but it does prove an interesting point,
events can be completely
deterministic as to source,
but completely random
for virtually all purposes
as to receiver.

viii) Speeding Up Bayesian Analysis: Ideal, Practical, or Qualitative

To speed any of these up,
choose your data D_e acquisitions
to be as decisive as possible.
To quote Ernest Rutherford (1871–1937)
"If you need statistics,
you did the wrong experiment"
(Trotta 2017, p. 4)
The pith of this aphorism is as above,
choose your data acquisitions
to be as decisive as possible.

17026

But Rutherford lived in simpler times where in physics at least, you need less statistics and, of course, Rutherford did use statistics as needed.

But in the modern age, in areas of cosmology, epidemiology, psychology, economics, social science, and AI statistics is what we've got.

But we have vast data sets and vast computing power make use of them.

Much of Bayesian analysis was worked out in the 1950s by Harold Jeffreys (no relation), but only with increasing computing power over the decades has Bayesian analysis grown in importance.

I think I'd never or barely heard of it before year 2000 and only since teaching cosmology I have become interested and NOT for practice, but just to understand why it's the path to truth (i.e., true theories)

ix) A key Point about Ideal Bayesian Analysis

You may start out very bad guesses at probable theories and your data acquisitions may be far less than decisive, but you will still approach truth if you keep doing the procedure (without making and repeating mistakes).

People often Bayesian analysis (or Bayesian Inference) is theoretically well justified and hopefully this section nonrigorously is a valid justification

People often say other statistical inference is NOT so well justified (but I have personal expertise on this point).

17.5 Bayesian Analysis: What People Actually Do Including Bayes Odds Ratios, and Bayes Factor But Deferring Marginalization to p. 170

So far as I know, no one in practical work does more than one explicit BA step of ideal BA in a paper. From paper to paper, it's just qualitative Bayesian analysis,

One-step Bayesian analysis

The ideal BA procedure described on p. 17011 – 17027, is just to show that ideal limit is true and you can approach that truth even just in one explicit Bayesian step.

Infact, what they do in that one-step BA is effectively to conflate multiple steps by expanding and contracting D_1 by including and excluding specific data sets.

A big practical problem is systematic error in data sets. By definition, the size of systematic error is unknown or you would be able to correct for it.

17029

Recent papers applying BA to cosmological data have spent a lot effort to detect systematic error by finding inconsistency between data sets.

Weird thing turn up: e.g.,

i) Two data sets support the same model but with inconsistent free parameters.

To some degree, this is the Hubble tension

ii) Two data sets support some of the same parameters, but for different models

I can't think of an example, but this may happen.

One thing people do NOT do is bother with the denominator $P(D_\ell | C_{\ell-1})$

$$= \sum_j P(D_\ell | T_j C_{\ell-1}) P(T_j | C_{\ell-1})$$

(see p. 17016)

It is easy to calculate given that you know the $P(T_i | C_{\ell-1})$'s and $P(D_\ell | C_{\ell-1})$'s, but you only need relative probabilities to judge the probability of truth (i.e., adequacy of theories in your set $\{T_i\}$.

(7030) This is true the ideal Bayesian analysis too. See p. 17017

What they do is compute the Bayesian posterior odds ratio or just the Bayes factor

(jargon of Kass & Raftery 1995, p. 776)

From p. 17017

$$OR_{post\,ij} \equiv \frac{P(T_i \mid C_\ell)}{P(T_j \mid C_\ell)} = \frac{P(D_\ell \mid T_i C_{\ell-1}) P(T_i \mid C_{\ell-1})}{P(D_\ell \mid T_j C_{\ell-1}) P(T_j \mid C_{\ell-1})} = K_{ij}^{(\ell)} * OR_{Pre\,ij}$$

$K_{ij}^{(\ell)}$ ← The (ℓ) means on step ℓ, not a power

Bayes K factor, but usage from Wik, NOT Jeffreys original definition (Kass, p. 776)

prior odds ratio

$$K_{ij}^{(\ell)} \equiv \frac{P(D_\ell \mid T_i\, C_{\ell-1})}{P(D_\ell \mid T_j\, C_{\ell-1})}, \quad OR_{Pre\,ij} \equiv \frac{P(T_i \mid C_{\ell-1})}{P(T_j \mid C_{\ell-1})}$$

$$K = \frac{\text{evidence for } T_i}{\text{evidence for } T_j}$$

evidence is synonym for marginal likelihood (Wik: Marginal likelihood)

Of course, in One-Step Bayesian analysis $\ell = 1$

17031

In One-Step BA, the initial probabilities $P(T_i|C_o)$ are usually just chosen by the principle of indifference because the set of theories $\{T_i\}$ has already been chosen by qualitative BA to be the liheliest theories to your knowledge.

With the choice $P(T_i|C_o)$ all equal, the relative judgment of adequacy of the theories all comes down to the Bayes factor K:

- the ratio of evidences (i.e., the ratio of marginal likelihoods)

which ratio or its logarithm!

I think the ratio of evidences sometimes is just called the "evidence" or the "Bayesian evidence" itself.

Jeffreys offered an Evidence Table judging the value of the evidence based on his examples or empirical testing (I think). Different versions have been offered, but I do NOT know of any optimum table.

17032

However, the Jeffreys table simplified by Kass & Raftery (1995, p.777) (probably because the original was speciously over-precise) is widely cited and may be the fiducial table.

Jeffreys Evidence Table for the Judgment of Bayesian Evidence (Bayesian K factor)

Simplified by Kass & Raftery (1995, p.777)

$K_{ij}^{(\ell)}$	$\log(K_{ij}^{(\ell)})$	Significance of Evidence
< 1	< 0	Negative (in favor of [redacted] T_j)
1 a 3.2	0 a 0.5	Insignificant
3.2 a 10	0.5 a 1	Significant
10 a 100	1 a 2	Strong
> 100	> 2	Decisive

in $K_{ij}^{(\ell)} = \frac{P(D_\ell | T_i C_{\ell-1})}{P(D_\ell | T_j C_{\ell-1})}$ where usually $C_{\ell-1} = C_0$ and $D_\ell = D_1$ for one-step Bayesian Analysis.

Note: Systematic errors could nullify the evidence

2026 jan 24 | 17033

Why is Evidence Table so cautious in being decisive?

Well, generally one is cautious about being decisive in science (e.g., the 5σ rule for claiming discovery). The evidence (using the term in its everyday sense) has to be strong because there can be so many mistakes in data and analysis. There's also Carl Sagan's (1939-1996) aphorism: "Extraordinary claims require extraodinary evidence" (ECREE), but others said similar things earlier (e.g., David Hume (1711-1776)).

However, to be specific to Bayesian analysis, in marginalization (see p. 170___), one must choose priors on the free parameters of theories and

7034

that choice is fraught with uncertainty and vast variation in what is thought to be good.

A poor choice of priors, can make a bad theory look good.

NOT priors on a theory but on the range of allowed for the free parameters for a theory in

marginalization

See p. 170

In fact, I think people are dismissive of $K_{ij}^{(\ell)} \leq 10$ and may NOT even consider $K_{ij}^{(\ell)} \simeq 100$ decisive.

But if $K_{ij}^{(\ell)} \gtrsim 1000$ that may be decisive even if the choice of priors is very poor.

Also a philosophical difference in choosing priors can arise

i) Choose them theory-independently

ii) Or choose prior (on the range of free parameters) that is physically plausible for a theory

~~4~~ 17.6

Multinomial Probability Distribution & The Multinonomid Theorem

2023dec05

~~23~~ 17035

1) There are probabilities, $P_1, P_2, P_3, \ldots P_I$ and $\Sigma P_i = 1$

Say you had I bins from which to draw events or in statistical mechanics terms in which to put classical particles

(classical but idonte

There probabilities are $P_1, P_2, \ldots$

row 1	P_1	P_2	P_3	P_4	$\ldots$	P_I
row 2	P_1	P_2	P_3	P_4	$\ldots$	P_I
row 3	P_1	P_2	P_3	P_4	$\ldots$	P_I
$\vdots$						
row N	P_1	P_2	P_3	P_4	$\ldots$	P_I

Do N selec wit repla in stats (Wi samp goin down the row

(Pointon-II calls these arrangements but those are equal probability in that book)

Then get sequences: e.g.

Sequence 1	$P_1 P_1 P_1 \ldots P_1$	$= P_1^N$	probability of this one sequ
seq 2	$P_1 P_2 P_1 \ldots P_1$	$= P_1^N P_2$	There are 2 sequen of equal probabil
seq 3	$P_2 P_1 P_1 \ldots P_1$	$= P_1^N P_2$	

in the sequence permuted the identical subscripts do not reveal new sequence

permuting the 1 and 2 indices of sequence 2 gives sequence which is different sequence

b) I have never found a good way in words for explaining why permuting identical indices gives NO distinct sequence. All I can say is look at the procedure on p. 17034. Note, that procedure is actually identical to expanding the multinomial $(P_1 + P_2 + \ldots + P_I)$ by power N; i.e.,

$(P_1 + P_2 + \ldots + P_I)^N$ which is why we call the Multinomial distribution the multinomial distribution. The expansion is the multinomial theorem.

The distinct sequences are called combinations (Wiki: Combination)

Combinations are identical if they conform to set $\{n_i\}$

where $\{n_i\}$ means

n_1 P_1's
n_2 P_2's
n_3 P_3's
$\vdots$
n_I P_I's

The order of the P_i does NOT cause non-identicallity.

In the jargon of statistical mechanics (at least according to Pointon-10-12), the set $\{n_i\}$ is called a configuration and the number of conforming combinations is weight $C(\{n_i\})$ of the configuration.

How does one calculate weight $C(\{n_i\})$ and the probability of selecting a conforming combination?

c) Weight $N! = C(\{n_i\}) \prod_i n_i!$

For example consider sequence of elements $P_1\ P_2\ P_1\ P_1\ P_6\ P_1 \ldots\ldots P_{I-3}$

N elements in ~~the~~ selection

$\therefore$ $N!$ permutations

i.e., N ways of selecting a first element,
$N-1$ ways of selecting a second element,
etc.,
but permuting the n_i P_i's among themselves creates NO new combination (i.e., distinct sequence).

Therefore, you can factorize $N!$ as above and get

$$C(\{n_i\}) = \frac{N!}{\prod_i n_i!}$$ which is the weight of the configuration $\{n_i\}$ and also the multinomial coefficient of the multinomial theorem.

Note, if there are just two elements P_1 and P_2, then
$n_2 = N - n_1$ and
$C(\{n_i\}) = \frac{N!}{n_1!(N-n_1)!} = \binom{N}{n_1}$ which is the binomial coefficient.

17038

d) Probability of Selecting a Combination Conforming to Configuration $\{n_i\}$

Now the P_i's are the probabilities of events i (see p. 17035).

So the probability of selecting a specific combination conforming to $\{n_i\}$ is $\prod_i P_i^{n_i}$

But what is the total probability of selecting any combination conforming to $\{n_i\}$ (i.e., the probability of getting configuration $\{n_i\}$)?

One guesses $P(\{n_i\}) = C(\{n_i\}) \prod_i P_i^{n_i}$,

but I can't see any proof just based on what we've said so far

However the probability of getting any combination conforming to any combination or in other words of getting any sequence is 1.

$$1 = 1^N = (P_1 + P_2 + P_3 + \dots P_I)^N = \left(\sum_i P_i\right)^N$$

When you actually do the multinomial expansion, it is clear (maybe) that the probability of configuration $\{n_i\}$ being obtained is $P(\{n_i\}) = C(\{n_i\}) \prod_i P_i^{n_i}$

All sequences produced uniquely, therefore all combinations with their correct weighting (see p. 17037)

2025dec05

There the count of combinations conforming to set $\{n_i\}$

$$C(\{n_i\}) = \frac{N!}{\prod_i n_i!}$$

which must be an integer

25
1703 9

The probability of combinations have equal value of ~~only one sequence~~ and this value ~~gives~~ $\{n_i\}$ is $\prod_i P_i^{n_i}$

Also probability of any sequence conforming to set $\{n_i\}$

∴ The probability of getting a combination conforming to set $\{n_i\}$ are ~~non-zero~~, ~~distribution~~ is

$$P(\{n_i\}) = C(\{n_i\}) \prod_i P_i^{n_i}$$

Probability distribution ~~of the distribution of the combinations~~ of ~~set~~ $\{n_i\}$

number of ~~sequence giving the distribution~~ combinations conforming to set $\{n_i\}$

Probability of getting any one ~~sequence~~ of the combination

$$= \frac{N!}{\prod_i n_i!} \prod_i P_i^{n_i}$$

e) Now $$P(\{n_i\}) = N! \prod_i \frac{P_i^{n_i}}{n_i!}$$

As we know from statistical mechanics this is the distribution for classical identical particles.

The probability distributions for bosons & fermions are different

In a QM sense, wavefunctions must obey the symmetrization principle. Boson/Fermion wavefunctions must be symmetric/antisymmetric

Bose-Einstein statistics

Fermi-Dirac statistics

we often say they are because the particles are identical but I think we just mean Quantum identical not classical identical

26
7040

— and the particle statistics exist from the symmetrisation principle in QM — which nature demands to prevent infinite degeneracy of multi-particle states or so it seems.

Normalisation?

Proof +

$$1 = \sum_i P_i$$ ← sequences of

$$1 = \left(\sum_i P_i\right)^2$$ = binomial theorem

$$1 = \left(\sum_i P_i\right)^N$$ = multinomial theorem

With polynomial one coefficient and variable ax^2+bx+c etc. (Wik)

we demand for probability particles add up to 1 — sequence of 2 sequences of N

f) If $i = 2$ you have the binomial theorem

$$(P_1 + P_2)^N = \sum_{R=0}^{N} \binom{N}{R} P_1^R P_2^{N-R}$$

Binomial coefficient

and

$$P(\Sigma n_i \xi) = \binom{N}{R} P_1^R P_2^{N-R}$$

for higher multinomials the probab. are harder to get formula easy formula for, Are no easy ones Elliot.

The key point is that in multiply out the expanding into and collect get just the sequences and collecting the coefficient the like terms, the multinomial probability distribution

e.g., $(P_1 + P_2)^2 = P_1 + P_1P_2 + P_2P_1 + P_2^2 = P_1^2 + 2P_1P_2 + P_2^2$

2025dec05

17041

Of course, if you actually started collecting sequences of putting N particles into I states, it is only in the limit of $I \to \infty$ sequences that you would recover the multinomial probability distribution

Omit: Digression

9) Further Digression on Binomial Theorem

Special case: Binomial theorem & Binomial probability distribution, Binomial Thm

Prove by induction if afflicted by paranoia.

$$1 = (p+q)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

for $q = 1-p$

$$= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$$

Moments of the binomial distribution.

A old trick

Define function

$$f(x) = (x+q)^n = \sum_{k=0}^{n} \binom{n}{k} x^k q^{n-k}$$

the n choose k formula

Just called binomial coefficient (Wilf)

$$M_\ell = \left(x \frac{d}{dx}\right)^\ell f \Big|_{x=p} = \left(x \frac{d}{dx}\right)^\ell (x+q)^n \Big|_{x=p}$$

$$\sum_{k=0}^{n} k^\ell \binom{n}{k} x^k q^{n-k} \Big|_{x=p}$$

8)

~~28~~
1702

$k=0 \qquad M_0 = 1$

$k=1 \qquad M_1 = x\, n (x+q)^{n-1} \Big|_{x=p} = np$

$k=2 \qquad M_2 = n x \frac{d}{dx}\left[x (x+q)^{n-1} \right] \Big|_{x=p}$

$= np \left[(x+q)^{n-1} + x(n-1)(x+q)^{n-2} \right)$ $\quad x = p$

$= np \left[1 + p(n-1) \right]$

$= pn + p^2 n (n-1)$

$\sigma^2 = \langle (x - \bar{x})^2 \rangle = \langle x^2 - 2x\bar{x} + \bar{x}^2 \rangle$

$= \langle x^2 \rangle - \bar{x}^2 = pn + p^2 n(n-1) - p^2 n^2$

$= p^2 n(n-1) \qquad = pn - p^2 n$

$= pn(1-p)$

BeV-53
сегмент

Probably now totally worthless

2021 nov 22

Theory 1

$P_i = \frac{1}{(\dots)}$ $\sum r_i = 1$ or 2 should

Throw sample N words and get $\{n_i\}$

$$P(\{n_i\}) = \frac{N!}{n_I! \dots n_2! n_1!} \prod_{i=1}^{I} \left(\frac{1}{I}\right)^{n_i} \qquad \sum n_i = N$$

$$= \frac{N!}{n_1! \dots n_I!} \left(\frac{1}{I}\right)^N$$

Theory 3

$$P(\{n_i\}) = \frac{N!}{n_1! \dots n_I!} \prod_{i=1}^{I} \left(\frac{i}{\frac{I(I+1)}{2}}\right)^{n_i}$$

$$= \frac{N!}{n_1! \dots n_I!} \frac{1}{\left(\frac{I(I+1)}{2}\right)^N} \prod_{i=1}^{I} i^{n_i}$$

As so often in Astronomy/God has told us there are only 2 possible true theories

Theory 2

$$P(\{n_i\}) = \frac{N!}{n_1! \dots n_2!} \left(\frac{1}{2I}\right)^N \prod_{i=1}^{I} [2 - \mathrm{mod}(i,2)]^{n_i}$$

Very easy to "postdict" $P(D|T)$ but like complex system with huge data

5) Die Problem for Zeus II

Normal problem

$I = 6$ Theory 1 $P_i = \frac{1}{6}$

Theory 2 $P_i = 2 - \mathrm{mod}(i,2)$ $\left\{ \begin{array}{l} \frac{1}{9} \text{ if } i \text{ odd} \\ \frac{2}{9} \text{ if } i \text{ even} \end{array} \right.$

$P(T_1) = \frac{1}{2}$

$P(T_2) = \frac{1}{2}$

Prior odds $\frac{P(T_1)}{P(T_2)} = 1$

Principle of indifference (Bayes p. 6)

So throw complex enough, but decisive?

$(N = 10)$

	1	2	3	4	5	6	7	8	9	10
	1	5	1	1	2	5	2	6	3	4

Really did this in 2019

$$P(D|T_1) = \frac{10!}{3!\,2!\,2!\,1!\,1!\,1!} \cdot \left(\frac{1}{6}\right)^{10} = 2.50057\ldots \times 10^{-3}$$

$$P(D|T_2) = \dots \left(\frac{1}{9}\right)^{10} 1^6 \cdot 2^4 = 6.938\ldots \times 10^{-4}$$

mat

$P(T_1|K_1) = \frac{P(D_1|T_1,K_0)}{P(D_1|T_1,K_0)P(T_1|K_0) + P(D_1|T_2)P(T_2|K_0)} * P(T_1|K_0)$

Probability given your knowledge before you've seen any experiments

$K = 3.604$

So Prior odds = 1

Bayes factor: $K = 3.604$

Posterior odds = 3.604

If there are two theories really

$P(1) = \frac{3.604}{4.604} = 0.7828...$, $P(2) = \frac{1}{4.604} = 0.21719...$

If one carried on the experiment to $N \to \infty$ for a die,

$P(1) \to 1$

$P(2) \to 0$

But my dice is not perfect and it can be loaded

Face 4 opens!!

I can load it. It's a road to Vegas

$P_1(1) = 0.7828$ $P_2(2) = .21719...$ do posteriors → priors

Do another set of 10

1	2	3	4	5	6	7	8	9	10
6	1	3	6	6	2	1	5	3	5

10.9.1.7.6.5.4

$P(D|T_1) = \frac{10!}{3!2!2!1!2!} \left(\frac{1}{6}\right)^{10}$

$P(D|T_2) = \ldots \left(\frac{1}{3}\right)^{10} \left(\frac{1}{6}\right)^{4} (2)^{4}$

The House never loses, And we are the House

17.7

a Toy problem

Die Problem Solved by Bayesian Analysis

One die 

Possible events $I = 6$

Theory 1 $P_i = \frac{1}{6}$, $\sum_i^I P_i = 1$

Theory 2 $P_i = \begin{cases} \frac{2 - \text{mod}(i,2)}{9} \\ \frac{1}{9} \quad i \text{ odd} \\ \frac{2}{9} \quad i \text{ even} \end{cases}$

$\sum_i P_i = 3(\frac{1}{9}) + 6(\frac{2}{9}) = 1$

Probabilities Normalized for 2 theories of how a die works on trials (i.e., throws of the die)

Prior probabilities

$P(T_1 | C_0) = \frac{1}{2}$

$P(T_2 | C_0) = \frac{1}{2}$

Using principle of indifference just Set them equal since we don't know which is true. But we could have assigned them any how and eventually the BA would have found truth.

So I did 10 throws (in 2019 actually)

Data = D_1

Throw	1	2	3	4	5	6	7	8	9	10
event	1	5	1	1	2	5	2	6	3	4

Calculated Marginal Likelihoods

likelihoods

$P(D_1 | T_1 C_0) = \frac{10!}{3!\,2!\,2!\,1!\,1!\,1!} (\frac{1}{6})^{10} = 2.4007\ldots \times 10^{-7}$

Need multinomial theorem in fact which there is a proof of in my notes (see p. 17035)

1's, 2's, 3's, 4's, 5's, 6's

the probability of each throw

The number of ways this sequence could happen counting over orders of sequences

$P(D_1 | T_2 C_0) = \frac{10!}{\ldots} (\frac{1}{9})^{10} \times (1)^4 (2)^4 = 6.738\ldots \times 10^{-}$

same

$= 86$ odd event even events

~~37~~ 17046

(2025 ~~May 01~~)

First Posterior odds ratio (OR_{Post})

$$\frac{P(T_1 \mid \mathcal{K}_1)}{P(T_2 \mid \mathcal{K}_1)} = \frac{P(D_1 \mid T_1\, \mathcal{K}_0)}{P(D_1 \mid T_2\, \mathcal{K}_0)} \frac{P(T_1 \mid \mathcal{K}_0)}{P(T_2 \mid \mathcal{K}_0)}$$

Prior odds ratio (OR_{Pre})

$\frac{P(T_1 \mid \mathcal{K}_0)}{P(T_2 \mid \mathcal{K}_0)} = 1$

Bayes Factor
Bayes K factor = K_{12}

~~But this not my background knowledge~~ $K_{\mathcal{K}}$

$$K_{12} = \frac{2.500577\ldots \times 10^{-3}}{6.938\ldots \times 10^{-4}}$$

$= 3.604$

Our Bayesian evidence or the $\log_{10}(K)$ is called the evidence.

which is also the Posterior odds ratio, given equal probabilities in prior odds ratio

See p. 17032 for a different version

Jeffrey's table based on some empirical ~~testing~~ (I think) (see p. 17032)

K	Evidence strength
< 1	negative support
1 - 3	worth mentioning
3 - 10	substantial
10 - 30	strong
30 - 100	very strong
> 100	decisive

Of course, this is just fiducial set of values and systematic errors could nullify the evidence.

In our case theory 1 is substantially favored over theory 2.

(2023 may 01)

17047

Most of the formal improvement from prior probabilities to posteriors

$$P(T_i \mid C_\ell) = \frac{P(D_\ell \mid T_i C_{\ell-1})}{\langle P(D_\ell \mid T_i C_{\ell-1}) \rangle} P(T_i \mid C_{\ell-1})$$

posterior — ratio of likelihood to average likelihood — prior

$$P(D_\ell \mid C_{\ell-1}) = \sum_j P(D_\ell \mid T_j C_{\ell-1}) P(T_j \mid C_{\ell-1})$$

Our best estimate of $P(D_\ell \mid C_{\ell-1})$

Probability of getting data D_ℓ given context C_ℓ (background knowledge).

An exactly true expansion if all our $P(D_\ell \mid T_j C_{\ell-1})$ and $P(T_j \mid C_{\ell-1})$ are exactly right and were averaging over all theories T_j with nonzero probability.

Note, Normalization

$$\sum_i P(T_i \mid C_\ell) = \frac{\langle \ldots \rangle}{\langle \ldots \rangle} = 1$$

and so even if the likelihoods and priors were only relative probabilities, the posteriors are normalized.

2025 May 01

$B_{12}^{(1)} = \dfrac{P(D_1|T_1, k_0)}{P(D_1|T_2, k_0)}$ — I divided these as a simplification

In our case

$$P(T_1|E_1) = \frac{(3.6\ldots)*\frac{1}{2}}{(3.6\ldots)*\frac{1}{2} + (1)*\frac{1}{2}}$$

$$= \frac{3.604}{4.604} \simeq 80\%$$

$$P(T_2|E_1) = \frac{1}{4.604} \approx 20\%$$

If my die were perfect (true) and not loaded, (it can be loaded)

$$\lim_{\ell\to\infty} P(T_1|E_\ell) \to 1$$

$$\lim_{\ell\to\infty} P(T_2|E_\ell) \to 0$$

If it is a real Vegas die. It can be loaded. First rule of gambling: the House never loses and we are the House

But I didn't believe this, so I did another data aquisition

Data D_2

throw	1	2	3	4	5	6	7	8	9	10
event	6	1	3	6	6	2	1	5	3	5

likelihoods

$$P(D_2|T_1 k_1) = \frac{10!}{3!\,2!\,2!\,1!\,2!}\left(\frac{1}{6}\right)^{10} = ?$$

(6's, 1's, 3's, 2's, 5's)

$$P(D_2|T_2 k_1) = \text{- - - - -}\left(\frac{1}{9}\right)^{10}(1)^{6}(2)^{4} = ?$$

(odd; even) — 16 just on p. 1

I didn't compute these

event 4 … on p. 17048, but that is common to both likelihoods

2025 may 01

$$\frac{P(T_1 \mid E_2)}{P(T_2 \mid E_2)} = B_{12}^{(2)} \frac{P(T_1 \mid E_1)}{P(T_2 \mid E_1)}$$

OR_{post}

$B_{12}^{(2)}$: accident/x the same

$\frac{P(T_1 \mid E_1)}{P(T_2 \mid E_1)}$: This is the previous OR_{post}

$OR_{pre} = K_{12}^{(1)}$

17049

$= 3.604\ldots$

$= B_{12}^{(2)} K_{12}^{(1)}$

$\cong 13$ strong evidence (see p 17044)

Formally this is posterior odds ratio, but the distinction to Bayes factor is I think not always m

$P(T_1 \mid E_2) = 0.9285\ldots$

$P(T_2 \mid E_2) = 0.07148\ldots$

I did calculate these probabilities

So theory 1 is looking strong and theory 2 is weak.

But those sequence of throw could accident have been unrepres ative o the true distributi So theory could still be tru

But if I just keep doing throw sequenc of 10 and only the probabilistic results.

Only ~~Even~~ the ~~sequences of 10 throws inf~~ would one theory go to Probality 1.

in the limit of infinite throws of 10

~~36~~
17050

But is there another way?

Remember Ernest Rutherford's Rule (p. 17025)

"If you need statistics, you've done the wrong experiment."

So instead of a sequence of throws,

I examine my die and see it has a 6-fold symmetry.

Now it's not perfect symmetry.

But do I care about my particular die?

No! I care about the ideal die.

The die all dice aspire to be.

The ideal die has exact 6-fold symmetry.

Note, this is why Laplace so often interpreted broadly

Data D_3 is exact, 6-fold symmetry

$$P(D_3 \mid T_1\ \mathcal{K}_2) = 1$$

$$P(D_3 \mid T_2\ \mathcal{K}_2) = 0$$

$$\mathcal{K}^{(3)}_{B2} = \frac{1}{0} = \infty$$

and $P(T_1 \mid \mathcal{K}_3) = \dfrac{P(D_3 \mid T_1, \mathcal{K}_2)\, P(T_1 \mid \mathcal{K}_2)}{P(D_3 \mid T_1, \mathcal{K}_2) P(T_1 \mid \mathcal{K}_2) + 0} = 1$

$$P(T_2 | \mathcal{K}_3) = 0$$

So the Bayesian Analysis of this trivial toy problem has found truth — about ideal dies.

~~But of what if you~~

In fact, this ideal BA with steps of data acquisition is not really ever done in practise — or very rarely for some special case

See discussion on p.17028

Rather people just collect a set of theories that they can assign equal probability by principle of indifference and then calculate the Bayes factor

$$R_{ij} = \frac{P(D_1 | T_i \mathcal{K}_0)}{P(D_1 | T_j \mathcal{K}_0)}$$

and they slot various sets of data or combinations of data in.

~~38~~
.7052

2025may01

In cosmology they try this combination of data and then that and try to figure out whether there are systematic errors in one set or another and do all kinds of finicky tests.

The problem is the Λ-CDM model works so well that even small systematic ~~errors~~ errors can make Bayesian analysis indeterminate in practice and that maybe where we are. plus other data (2025)

In 2025 DESI DR2 came out. In 2026 April, Λ-CDM is still viable in the view of many

Does DESI DR2 rule against Λ-CDM by 3σ (which is BA evidence).

Yes — if the systematics are under control. But some argue they are NOT.

But also BA in real cases requires Marginalization.

17.8

Marginalization

2025 may 01

~~29~~ 17053

a) Most theories Bayesian Analysis is applied to have free parameters that must be set ~~by the~~ data itself

data interpreted broadly → it could a theory about some other aspect of reality than the ~~aspect~~ you are studying.

Now a theory with free parameters could be regarded as an infinite continuum of theories,

But that is not a useful perspective,

It's better to treat theories with free parameters as different from discrete theories and needing a different treatment to a degree.

Say you have a set of theories $T_i(\theta)$

θ is just a symbol for the set of free parameters which can be many and can differ between the theories, but we are being general and just need a symbol

All of which are continuous variables (i.e. real numbers or complex numbers)

which perhaps are almost always ~~to~~ except for trivial/toy problems like the die problem

~~p. 1704~~

17054 | 2026 jan 24

b) Marginalization

According to Wikipedia (Marginal distribution) "marginalization" comes from summing entries in a table column in a margin of the table. Marginalization meant using the sums rather than the entries.

For use in Bayesian analysis, "integration" might be more descriptive of what is done, but "integration" may be too general to give the special use.

Recall from p. 17017

$$P(T_i \mid C_\ell) = \frac{P(D_\ell \mid T_i C_{\ell-1})\, P(T_i \mid C_{\ell-1})}{\sum_j P(D_\ell \mid T_j C_{\ell-1})\, P(T_j \mid C_{\ell-1})}$$

$$= \frac{P(D_\ell \mid T_i C_{\ell-1})}{\langle P(D_\ell \mid T_j C_{\ell-1}) \rangle} P(T_i \mid C_{\ell-1})$$

↳ Recall the denominator is usually only of formal interest (see p. 17017)

Now to go "continuum family of theories" which now call just one theory with free parameters which are usually real or complex number variables, we expand

$$\underbrace{P(D_\ell \mid T_i\, C_{\ell-1})}_{\text{marginal likelihood or evidence}} = \int \underbrace{P(D_\ell \mid T_i(\theta)\, C_{\ell-1})}_{\text{likelihood = probability of } D_\ell \text{ given } T_i(\theta) \text{ and } C_{\ell-1}} \underbrace{P(\theta \mid T_i\, C_{\ell-1})}_{\text{probability density of } \theta \text{ given } T_i \text{ and } C_{\ell-1}} d\theta$$

2026jan24 | 17055

Actually, my basic reference (Kass & Raftery 1995, p. 776) ambiguously refer to both $P(D_e|T_i(\theta)C_{e-1})$ and $P(\theta|T_iC_{e-1})$ as "densities". Maybe there is choice on which is a probability and which is a probability density.

But the product $P(D_e|T_i(\theta)C_{e-1})P(\theta|T_iC_{e-1})$ is clearly a probability density.

In any case, I like to think of likelihood $P(D_e|T_i(\theta)C_{e-1})$ as a probability NOT a probability density.

If $T_i(\theta)$ is an exactly deterministic theory with exact setting (initial conditions not included in parameters) and D_e is exact (ie., error free), then $P(D_e|T_i(\theta)C_{e-1}) = \begin{cases} 1 & \text{which it may do for multiple } \theta \text{ and ranges of } \theta, \text{ which doesn't mean } T_i(\theta) \text{ is true just that it is adequate for the dat} \\ 0 & \text{otherwise} \end{cases}$

However, this specification is an over-idealization for practice because,

(i) D_e will be drawn from some uncertainty distribution,

(ii) $T_i(\theta)$ may have some uncertainty in initial conditions,

and (iii) many, maybe most,

17056 | interesting theories for Bayesian analysis will be probabilistic at least for reason (ii)

→ This is true for cosmology where the initial density fluctuations (i.e., primordial density fluctuations (Wik)) are randomly determined by theory and a random selection is used to initial large-scale structure formation.

Thus, in practice must expand likelihood in an integral

$$P(D_e|T_i(\theta)C_{e-1}) = \int \underbrace{\mathcal{L}_{\text{likelihood}}(\ldots)}_{\text{a likelihood probability density}} \ldots$$

Then $P(D_e|T_i(\theta)C_{e-1})$ will range over $[0, 1]$

Calculating $P(D_e|T_i(\theta)C_{e-1})$ for interesting cases is a major chore (i.e., bore).

D_e can be petabytes (e.g., DESI final data will be a petabyte (10^{15} bytes) (Google AI))

and probably you have always forward calculate from $T(\theta_i) C_{e-1}$ to D_e. I did a reverse calculation for my die example (see p. 17045), but that was a very easy case.

One often uses Markov chain Monte Carlo (MCMC) to predict data as a function of inputs. A big calculation for big data.

Yours truly knows little of MCMC though I'm somewhat knowledgeable about Monte Carlo radiative transfer.

But after that you still have to calculate the marginal likelihood

$$P(D_e | T_i C_{e-1})$$

which usually requires another huge multi-dimensional integral weighted by the prior probability density for free parameters

$$P(\theta | T_i C_{e-1}).$$

So overall calculating $P(D_e | T_i C_{e-1})$ is a huge computation for interesting cases.

Key point: you are marginalizing (i.e., integrating over ranges of free parameters! i.e., the priors).

If you know a theory is true (i.e., fully adequate), then you maximize the likelihood to find the best parameters (see p. 170 ___).

But here BA is used to try to find the most adequate theory without fine-tuning parameters to fit the theory to particular data sets (where you could just be fitting noise or systematic error).

2026jan24 (17052

If the Bayesian evidence ratios
favor a particular theory
by huge factors (e.g., $\gtrsim 100$)
for plausible priors, then
that theory is,
the most adequate
no matter what free parameters
are chosen for each theory
and may be the true theory.

Marginalization implements
Occam's razor mathematically
in that it does NOT
favor theories with many free parameters
just because they can
be tightly fit to particular datasets
using Maximum likelihood.

40) 2024 may 01

b) A word about likelihoods with free parameters

$$P(D_e \mid T_i(\theta)\, K_{e-1})$$

- D_e: New data
- T_i: theory T_i with its set of free parameters
- K_{e-1}: Background knowledge which is usually left implicit

edo
124
of
elihood
IT probability
it it is!!
come
om Bayes
m.
epl2
t is NOT
probality
lensity

What if you know or assume theory $T_i(\theta)$ is true, but you don't know its parameter values

See p 49 for more on maximum likelihood

The general rule is Maximize the likelihood

Symbolically

$$\frac{\partial P(D_e \mid T_i(\theta)\, K_{e-1})}{\partial \theta} = 0$$

and solve for set θ

With potentially of data many parameters this may be a super-computer calculation eg. stochastic gradient descent

Remember the theory may just be statistical and also the data D_e may statistical error

but we assume no systematic error (but in practice there can be)

But because of the statistical aspect the D_e may be unrepresentative of the true parameter values

So you can get very bad values

But if $D_e \rightarrow$ {large / decisive} $\rightarrow$ infinite

then eventually you will get the true parameter values.

But again if your theory is not right the parameter values may be wrong or meaningless in that they may NOT be the true parameters.

Also if your theory has many parameters, you may simply get an excellent fit to a set of data and excellence argues for it being right $\rightarrow$ but this may wrong $\rightarrow$ you may be fitting noise for example.

42

marginalization ideally doesn't worry about ~~best~~ fitting parameters, but about the truth of theories, in fact, marginalization is marginalizing over parameter values

In a real way, marginalization embodies Occam's razor putting theories on an equal basis no matter how many free parameters they have.

c) Marginalization

For theories $T_i(\theta)$, instead of prior probabilities you have prior probability densities.

$$dP(T_i \mid k_{e-1}) = P(T_i(\theta) \mid k_{e-1})\, d\theta$$

Despite what Laredo 2024 may say, they misspeak

The likelihood is still a probability NOT a probability density

differential symbolizing all parameters

$$dP(T_i \mid k_e) = P(T_i(\theta) \mid k_e)\, d\theta$$

$$= \frac{P(D_e \mid T_i(\theta) k_{e-1})\, P(T_i(\theta) \mid k_{e-1})\, d\theta}{\sum_{jm} P(D_e \mid T_i(\theta_m) k_{e-1})\, P(T_j(\theta_m) \mid k_{e-1})\, d\theta_m}$$

The sum on m can just be turned in to integrals

$$P(T_i(\theta)|K_e)\,d\theta = \frac{P(D_e|T_i(\theta)K_{e-1})\,P(T_i(\theta)|K_{e-1})\,d\theta}{\sum_j \int P(D_e|T_j(\theta)K_{e-1})\,P(T_j(\theta)|K_e)\,d\theta}$$

$\sum_j$ ← mean likelihood again.

Different integrand but same economized on symbols and indices

So in an idealized BA, one could ~~get~~ ~~record~~ constructed updated probability densities in a series of steps

but that's probably never done.

One just integrated over θ

$P(T_i(\theta)|K_e)$ [illegible]

$$P(T_i|K_e) = \frac{\int P(D_e|T_i(\theta)K_{e-1})\,P(T_i(\theta)|K_e)\,d\theta}{\langle \text{mean likelihood} \rangle}$$

θ integrated away

$\sum P(T_i|K_e) = 1$ 'Normalized automatically'

in fact, the series stops usually at $l = 1$. (always)

94) And who cares about the denominator. We just want the Bayesian evidence

$$B_{ij} = \frac{P(T_i \mid K_e)}{P(T_j \mid K_e)} = \frac{\int P(D_e \mid T_i(\theta) K_e) P(T_i(\theta) \mid K_e) d\theta}{\text{same for } T_j}$$

integrate away → marginalize the parameter

Really this Bayes posterior odds ratio is some ~~the~~ old-fashioned ~~formula~~ score (Kass & Raftery 1995)

but I think everyone just calls it the Bayes factor or Bayes evidence or the log of it the Bayes evidence.

d) Pavone

What is the difficulty since everything looks simple formally.

What is $P(T_i(\theta) \mid K_{e-1})$?

usually ~~$e-1=0$~~ $e-1=0$

What are the priors ?

The commonest approach is the flat prion

$$P(T_i(\theta)|k_{i-1}) = \begin{cases} \frac{1}{\theta_2-\theta_1} & \text{for } \theta \in [\theta_1, \theta_2] \\ 0 & \text{otherwise} \end{cases}$$

You can do more tricky distributions but You'd have to have a good reason

$\theta_1 \Rightarrow$ best guess at absolute lowest value θ could be

$\theta_2 \Rightarrow$ best guess at absolute highest θ could be.

In fact, people estimate θ_1 and θ_2 in all kinds of ways

It is the ART of Bayesian analysis

Choice is Qualitative Bayesian analysis

46

Ok, you can do that
But what's the problem
then?

Doing the integral

$$\int P(D_e \mid T_i(\theta) K_{e-1}) P(T_i(\theta) \mid K_{e-1}) d\theta$$

You have to predict for maybe
~~maybe~~ petabytes of ~~data~~ data
the likelihood
from a multidimensional
theory and then do a multidimensional integral
→ It was easy for
the die problem (see p-31),
but it's not easy
for the DESI DR2.

Use advanced multi-dimensional
integral methods
— probably Markov Chain Monte Carlo
(with which I have no experience
though I know something
about Monte Carlo Radiative
transfer)

So, in fact, state of art Bayesian analyses on huge data is ~~enormous~~ computationally

Probably, Machine learning is helping.

Also there are ~~simplex~~ quick methods of estimating Bayesian evidence: e.g.

BIC = Bayesian information criterion

AIC = Akaike Information criterion

which I think well approximate Bayesian evidence in certain useful limits

but I know little of them.

e) ~~BEST~~ The Bayesian Evidence

$$B_{ij} = \frac{P(T_i|K_e)}{P(T_j|K_e)}$$ see p. 44

You've marginalized over the parameters. The evidence is based on how good the theory is no matter how many parameters there are and (if your priors are good) no matter what the ~~best~~ parameter values are.

48

Because of the slop
in the art of priors
and systematic errors,
No one is impressed
by evidence $B_{ij} \leq 10$
maybe not even $B = 100$,
but if it is $B_{ij} = 1000$,
then theory T_i is probably
a lot better than T_j
and T_j is probably
ruled out.

f) Maximum Likelihood Redux

What is it from a Bayesian analysis perspective? A mythical interpreted discussion

Recall from p. 43

$P(T_i(\theta)|K_e)d\theta = P(D_e|T_i(\theta)|K_{e-1})P(T_i(\theta)|K_{e-1})$

mean likelihood

$\sum_i \int P(T_i(\theta)|K_e)\,d\theta = 1$ normalized automatically

But say you take T_i as true and just normalize

$1 = \int P(T(\theta)|K_e)\,d\theta = A\int P(D_e|T_i(\theta)|K_{e-1}) \times P(T_i(\theta)|K_{e-1})d$

normalization constant

Then you could read

$P(\theta) = P(T(\theta_i)|K_e) = A\,P(D_e|T_i(\theta)|K_{e-1})P(T_i(\theta)|K$

as probability density for obtaining parameter values θ.

~~for obtain parameter value~~

~~Regard D_e as a measurement of θ~~

50 | But what is the measurement for obtaining the actual θ_{true}?

Obtaining ~~all~~ K_{emax}

All knowledge relevant to the theory,

But what does the probability density describe?

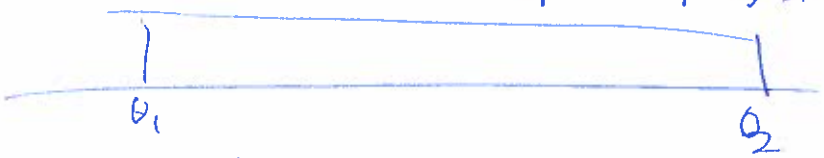
Maybe an ensemble of universes where T_i is true in ~~everybody~~ all of them, but with a continuum of varying θ,

You don't know what universe you are in until you get K_{emax}

$P(\theta)d\theta$ is probability for ~~the~~ your universe being ~~in~~ in $d\theta$.

~~Posterior~~ probability density

Assume $P(T_i(\theta) | K_{bkd})$ is flat between known bounds on θ as a ~~step~~ sweeping simplification

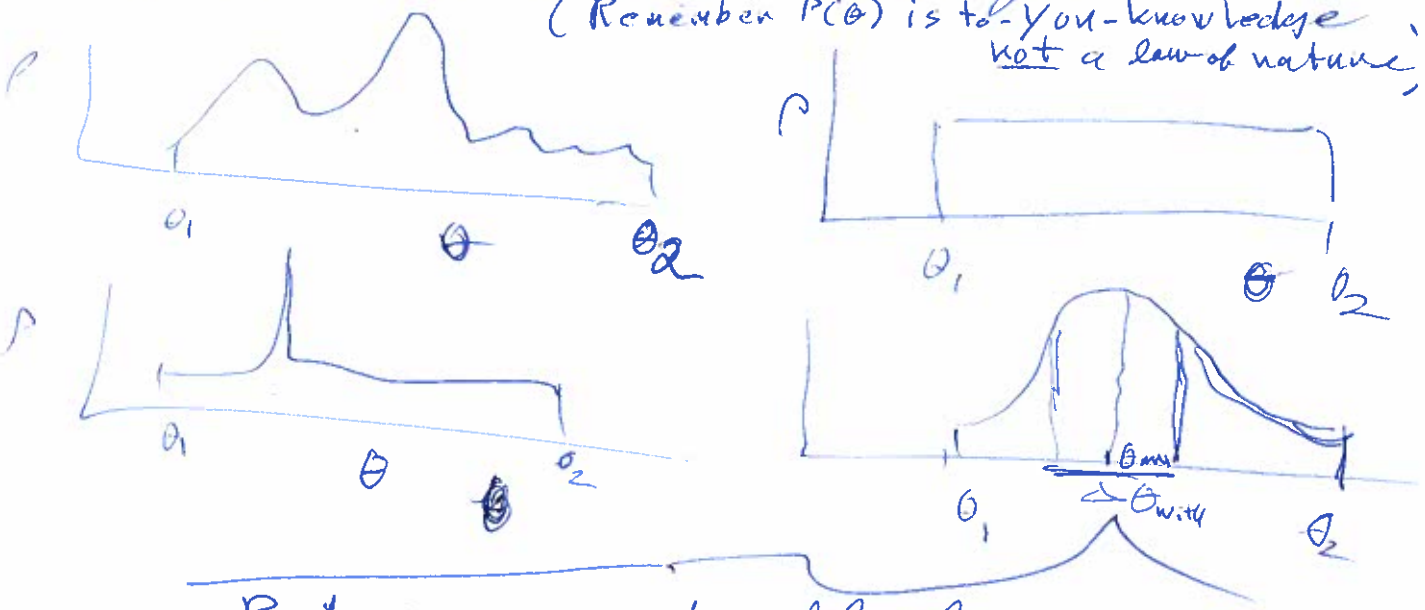


$\therefore P(\theta) \propto P(D_e | T_i(\theta) K_{e-1})$

2025 may 01

$P(\theta)$ could be anything in general

(Remember $P(\theta)$ is to-your-knowledge not a law of nature)



But suppose it is like this.

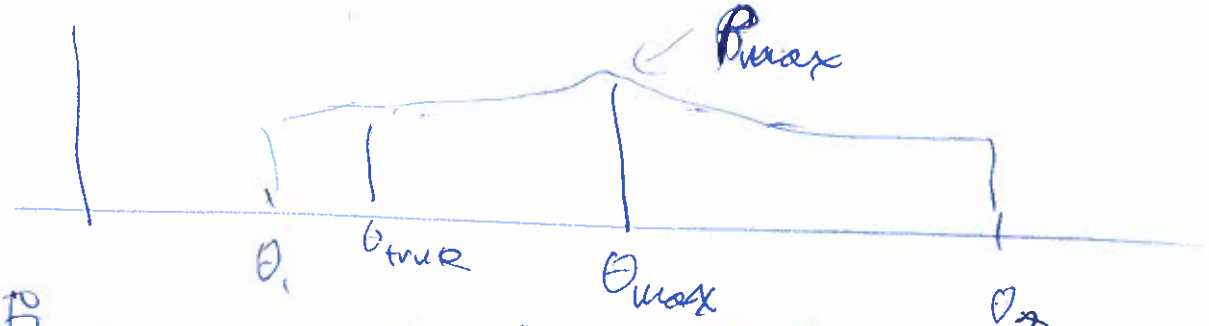
Most probability is clustered within $\Delta\theta_{width}$ of θ_{max}.

$\therefore$ $[\theta_{max} - \Delta\theta_w, \theta_{max} + \Delta\theta_w]$

Say $P(\theta_{max}, \Delta\theta) = 0.9$ or 0.99 or 0.999

$\therefore$ θ_{max} is a good approx to θ_{true} if $\Delta\theta_w$ is rather small.

But $P(\theta)$ may not be so favorable.

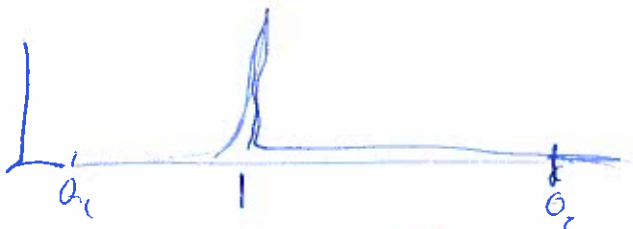


θ_{max} may be far off from θ_{true}

So if $P(\theta)$ is good
θ_{max} should be close
to θ_{true},
but it can be very bad.

How do you make $P(\theta)$ good?
find D_ℓ decisive
following Rutherford rule

$$P(D_\ell \mid T_i(\theta), \alpha_\ell)$$



θ_{max} of a near Dirac delta function.

– almost only the true θ can give data D_ℓ.

And if your theory isn't true?
You may never find decisive D_ℓ
But you can always try theory T_j

2019 nov 26

53

If you have vast sets
of data

Cosmological

peta bytes

1 pb = 10^{15} bytes

LHC in 2015 pb per year

~~couple~~ process

24 pb per day
in 2009

Integration
can be huge even
for our modern standard
just doing ~~it~~ follow

Are shortcuts of
varying utility of
which I no for know
little

a) Pangogial Information = DIC
eg Schwarz crit

Finding
likelihood
PCP | $T_i(\theta_i)$

for the
problem
simple
but
petabytes
produce
a
compl
then

I
intuit
that
here
you
need
Markov
Chain
Monte
Carlo
MCMC

Cotta-35
and a lot
skill

done
from
Bayes
an approx

(2019 nov 26)

derive from information theory

Akaike Information crit. (Wiki)

AIC

↳ relative quality of model

Not absolute

(from information theory)

$$AIC = 2k - 2\ln L$$

↑ k ↑ AIC

↑ L ↓ AIC

k: number of estimated parameters

max likelihood of model

$\ln\left(\frac{L_{max}}{L_{max}, AIC \text{ in min}}\right) \approx k$ or better odds

[AIC − AIC]/2 ≈ like k or better odds

lower 1, better

BIC NOT

similarly to

some different weightings

somehow related

BIC & AIC must work well in some limits & so used but can't be universally use or people