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

Bayesian Analysis (BA)

(A name I prefer to the usual Bayesian Inference)

1) Preamble: BA is path to truth generalized 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) Truth 6) Bayes Odds Ratio & Bayes Factor (p. 21)

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 Criterion
AIC = Akaike Information Criterion

~~Q~~
~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

I argue ~~that~~ Bayesian Analysis (BA = AkA Bayesian inference or Bayesian statistics) but done on computer packages

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~~ important theories are like their.

at least improvement. The Scientific Method Quantified and thereby the proof of the Sci. Method

E.g., Newtonian Physics.

↪ It is believed to be ~~the~~ macroscopic, limit of QM low velocity ~~limit~~ of ~~Special~~ relativity weak gravity limit of General relativity

In fact virtually all, Except maybe TOE (theory of everything or 2nd law of Thermodynamics) ~~But~~ are

Another ~~One~~ 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.

Maybe actually said? Well the system must be big enough?

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

Natural selection evolution

↳ unlike Newtonian Physics ↳ these can be proven by mathematical logic (and how? Well for any toy universe you imagine) — and so can Bayesian Analysis

Aside from pure rigorosity or pure philosophical gubbleity 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 thing will happen in a given situation

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

17004

These are your prior probabilities or priors

Based on vague approximate frequency $\frac{N_{event}}{N_{trial}}$ often with a lot of mistakes and gullibility

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

Your updating 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 injections can change the rules of success suddenly → (who 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.

~~6~~ 17006

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

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

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

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

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

Cont'd 2022 p. 3 tells me K is called conditioning information or context in BA jargon

Consider 3 events A, B, K.

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

I use ABK as joint event as a shorthand for union = ∪ symbol in math which is too klutzy for me in this context

order has no meaning.

2025 May 01

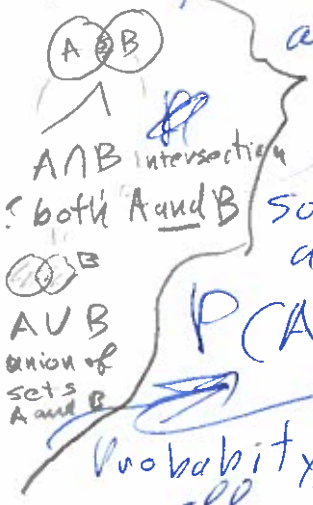
17007

Conditional Probability and Probability Product Rule

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

ABK is ~~conjoint~~ joint event, intersection all 3 at once

Background knowledge that in BA jargon is often called conditioning knowledge or context. (Loredo 2024 p 3)



ABK = ~~A∩B∩K~~ A∩B∩K

~~∩~~ symbol, intersection

but too klutzy for a shorthand and looks ugly in general

so NOT a product

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

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

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

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. (Twitt 2017 p 5)

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

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

17008

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

Root of Bayesian analysis

↳ Bayesian analysis was finally elaborated in 1990 ~~ON~~.
(Scientific Inference 3rd ed 1973
by Harold Jeffreys (1891–1989)
(no relation — but he managed 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 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.]

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

Actually, a actual practice. the ideal steps beyond step 1 are not done. Everyone seems to start step 1 with the hope that step 1's will gradually get better

they are set of done in single papers. Now we add this data, now this and take away that'

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

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 ~~on~~ 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, I say it just true or false

~~10~~
17010

speaking in an absolute sense not ~~wrong~~ having aside partially true 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 ~~of~~ 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 Restrid thein 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 philosopheus of science ave great quibblevs.

The proof is NOT absolutely vigorous 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

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 $\xi T_i \xi$'s, 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 those, 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

$P(T_i | C_o)$ conditional probability of T_i given context C_o

theory T_i ↑ ↖ 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 $P(T_i | C_o)$ equal.

You can normalize the probabilities if you like!

$$\sum_i P(T_i | C_o) = 1$$

but this isn't necessary since ideal BA normalizes updated theory probabilities automatically, (see p. 170(7))

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 (Wik)

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

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

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

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

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

$P(C_{\ell-1})$ 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

Actually, $P(D_\ell | T_i C_{\ell-1})$ 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 $P(D_\ell | T_i C_{\ell-1})$ (see p. 17...) we call it that here following the jargon of Kass & Raftery (1995, p. 776).

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

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

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

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

But this is 1 since we know it.

Really just a best estimate to your knowledge. You do NOT know $P(D_\ell | C_{\ell-1})$ in an absolute sense. You could include theories NOT in your set $\{T_i\}$, but that seems pointless practically AND as a formalism

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

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

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

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

Note the $P(T_i|C_\ell)$ are automatically normalized as noted on p. 17018: $\sum_i P(T_i|C_\ell) = 1$.

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.

Weighted mean of the marginal likelihood

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

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_ℓ continue, the probabilities of the theories will continue be updated.

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 $P(T_i|C_\ell) = 0$,

ii) However, unlucky sets of data D_ℓ may cause $P(T_{i*}|C_\ell)$ to decrease sometimes for some steps.

some may have reach $P(T_i|C_\ell)=0$ 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 re'assignment 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.

vi) Bad 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.

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 transformations.

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

17022

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 $<< C$, size scale >> microscopic, the local gravity field is sufficiently uniform, and one avoids some tricky electromagnetic effects.

It's a tune on 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 a series of 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 randomly 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, 17025
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"
(TvoHo 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 Boyes Odds Ratios, And Baxes Factor But Defering Marginalization to p. 17053

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.

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 $\{T_i\}$.

(17030) 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

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

$$K_{ij} = \frac{\text{evidence for } T_i}{\text{evidence for } T_j}$$

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 (Evidence) 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)

$K_{ij}^{(e)}$	$\log(K_{ij}^{(e)})$	Significance of Evidence
< 1	< 0	Negative (in favor of 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

---

i) $K_{ij}^{(e)} = \dfrac{P(D_e|T_i\, C_{e-1})}{P(D_e|T_j\, C_{e-1})}$ where usually $C_{e-1} = C_0$

and $D_e = D_1$

for one-step Bayesian Analysis

---

ii) Note: Systematic errors could nullify the evidence

2026jan24 | 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 similiar things earlier (e.g., David Hume (1711-1776)).

However, to be specific to Bayesian analysis, in marginalization (see p. 17053), one must choose priors on the free parameters of theories and

17034

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

In fact, I think people are dismissive of $K_{ij}^{(\ell)} \lesssim 10$ 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.

marginalization

See p. 170

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

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

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

~~There are probabilities $P_1, P_2, \ldots P_I$~~

row 1	P_1	P_2	P_3	P_4	...	P_I
row 2	P_1	P_2	P_3	P_4	...	P_I
row 3	P_1	P_2	P_3	P_4	...	P_I
⋮						
row N	P_1	P_2	P_3	P_4	...	P_I

Do N selection with replacement in stats jargon (W.k. sampling going down the rows

(Pointon-11 calls these arrangements but those are equal probability in that book)

You get sequences: e.g.

Sequ 1	$P_1 P_1 P_1 \ldots P_1$	$= P_1^N$	probability of this one sequence
seq 2	$P_1 P_2 P_1 \ldots P_1$	$= P_1^{N-1} P_2$	There are 2 sequences of equal probability
seq 3	$P_2 P_1 P_1 \ldots P_1$	$= P_1^{N-1} P_2$	

the identical subscripts does not give a new sequence

permuting the 2 and 2 indices of sequence 2 gives sequence 3 which is different sequence. (also a different distinct 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.

2026 Jan 24

17037

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: $P_1\ P_2\ P_1\ P_I\ P_6\ P_1 \ldots P_{I-3}$

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

~~25~~ 7039

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

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

which must be an integer

The probability of combinations have equal value of ~~only the 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 ~~the~~ combination conforming to set $\{n_i\}$ are ~~equivalent~~, ~~distribution~~ is

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

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

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

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

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

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 ~~&~~ fermions are different bosons

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
40

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

Normalization?

Proof

$1 = \sum_i P_i$ ← sequences of 1 which we demand for putting particles into box

$1 = (\sum_i P_i)^2$ — sequences of 2 = binomial theorem

$1 = (\sum_i P_i)^N$ sequences of N = multinomial theorem

Wiki polynomial one coefficient and variables ax^2+bx+c etc. (Wiki)

If $i = 2$ you have the binomial theorem

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

Binomial coefficient

and $P(\{n_i\}) = \binom{N}{\ell} P_1^{\ell} P_2^{N-\ell}$

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

The key point is that in ~~multiply out the~~ expanding ~~into and collect~~ you get just the sequences ~~and~~ 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 $l \to \infty$ sequences that you would recover the multinomial probability distribution

Omit; Digression

9) Further Digression on Binomial Theorem

Binomial Thm

Special case: Binomial theorem & Binomial probability distribution,

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 cool 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)

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

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

28) 17042

$\ell = 0$ $\quad M_0 = 1$

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

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

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

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

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

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

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

$= pn - p^2 n$

$= pn(1-p)$

BeV-53

Probably
Now totally
Worthless

2021 nov 22

17093

$p_i = \frac{1}{(\ldots)}$ $\sum r_i = 1$ or 2 should

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

Theory 1

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

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

This is for I words I runs $p_i = 1$ i odd $p_i = 2$ i even

I Total $= \frac{I}{2} \cdot 1 + \frac{I}{2} \cdot 2 = \frac{3}{2} I$

Theory 2

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

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

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

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

Very easy to ⇒ post dict! P(D|T) not like Taylor series with huge data

5) Die Problem for Zeus II

Normal problem $I = 6$

Theory 1 $p_i = \frac{1}{6}$

Theory 2 $p_i = \frac{2 - \mathrm{mod}(i,2)}{9}$ $\frac{1}{9}$ i odd, $\frac{2}{9}$ i even

$P(T_1) = \frac{1}{2}$ $P(T_2) = \frac{1}{2}$

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

Principle of indifference (Bayes's p. 6)

So throw complex enough, but not decisive?

$(N = 10)$

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

Really did this in 2019

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

$P(D|T_2) = \frac{10!}{3!\,2!\,2!\,1!\,1!\,1!} \cdot \left(\frac{1}{9}\right)^{10} = 6.17\ldots \times 10^{-4} = 6.938\ldots \times 10^{-4}$

Maths

6 i's and 4 2's

$P(T_1|K_0) = \frac{P(D_1|T_1 K_0) \times P(T_1|K_0)}{P(D_1|T_1 K_0) P(T_1|K_0) + P(D|T_2) P(T_2|K_0)}$

K = 3.604

Probability given your knowledge before even. No assumption

So
Prior odds = 1
Bayes K factor } k = 3.604
Posterior odds = 3.604

If these two theories really are exhaustive, one can calculate their probabilities updated ??

See p. 14 for update formula. Both updated & no... normalization factor

$P(1) = \frac{3.604}{4.604} = 0.7828$, $P(2) = \frac{1}{4.604} = 0.21799$

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

$P(1) \to 1$
$P(2) \to 0$

But my die is not perfect and it can be loaded Face 4 opens!!

I can load it. It's a real vegas die

$P_1(1) = 0.7828$, $P_2(2) = .21799$. . the posteriors → priors

Do another set of 10

10·9·1·7·6·5·4

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

The House never loses, And we are the House

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

$P(D|T_2) = \left(\frac{1}{5}\right)^{10} (1)^{6} (2)^{4}$

17.7

[2025 may 01]

[17045]

[31]

a Toy problem

Ideal

Die Problem solved by Bayesian Analysis (BA)

One die 

Possible events $I = 6$

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

Theory 2 $P_i = \frac{2 - \text{mod}(i, 2)}{9}$

$\frac{1}{9}$ i odd

$\frac{2}{9}$ i even

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

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

marginal likelihoods

$P(D_1 | T_1 C_0) = \frac{10!}{3!\,2!\,2!\,1!\,1!\,1!} (\frac{1}{6})^{10} = 2.5005\ldots \times 10^{-3}$

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 summing over orders of sequences

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

same

= 8/1 odd event : even events

~~17046~~ (2025 ~~May 01~~)

First Posterior odds ratio (OR_{Post})

$$\frac{P(T_1 \mid \cancel{E_1})}{P(T_2 \mid \cancel{E_2})} = \frac{P(D_1 \mid T_1 \cancel{E_0})}{P(D_2 \mid T_2 \cancel{E_0})} \underbrace{\frac{P(T_1 \mid \cancel{E_0})}{P(T_2 \mid \cancel{E_0})}}_{1}$$

Prior odds ratio (OR_{Pre})

Bayes Factor
Bayes K factor = $\cancel{B}_{12}$

~~But this not~~ ~~background knowledge~~ K_2

$$\cancel{B}_{12} = \frac{2.50057\ldots \times 10^{-3}}{6.938\ldots \times 10^{-4}}$$

$= 3.604$

Or 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 Evidence 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.

(2029 may 01)

17047

What of the formal improvement from prior probabilities to posteriors

$$P(T_i | C_e) = \frac{P(D_e | T_i C_{e-1})}{\langle P(D_e | T_i C_{e-1}) \rangle} P(T_i | C_{e-1})$$

Posterior

ratio of likelihood (marginal) to average marginal likelihood

prior

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

Our best estimate of $P(D_e | C_{e-1})$

Probability of getting data D_e given context C_{e-1} (background knowledge).

An exactly true expansion if all our $P(D_e | T_j C_{e-1})$ and $P(T_j | C_{e-1})$ are exacty right and were are averaging over all theories T_j with nonzero probability.

Note, Normalization

$$\sum_i P(T_i | C_e) = \frac{\langle \ldots \rangle}{\langle \ldots \rangle} = 1$$

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

~~34~~ 17048

2025 May 01

$$B_{12}^{(1)} = \frac{P(D_1|T_1 k_0)}{P(D_1|T_2 k_0)}$$

In our case

$$P(T_1|E_1) = \frac{(3.6\ldots)*\frac{1}{2}}{(3.6\ldots)*\frac{1}{2} + (1)*\frac{1}{2}} = \frac{3.604}{4.604} \approx 80\%$$

I divided through as a simplification

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

If my die were perfect (~~not~~ true) and not loaded,

It can be loaded

$$\lim_{e\to\infty} P(T_1|E_e) \to 1$$

$$\lim_{e\to\infty} P(T_2|E_e) \to 0$$

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 acquisition

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

marginal likelihoods

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

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

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

No event of 4

didn't compute these

every $\frac{1}{16}$ [illegible]

2025 may 01

17049

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

OR_{post}

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

accidental ≈ the same

This is the previous $OR_{post} = 3.609...$

$3.604... = K_{12}^{(2)} K_{12}^{(1)}$

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

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

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

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

I did calculate these probabilities

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

But these sequence of throws could accidentally have been unrepresentative of the true distribution. So theory 2 could still be true

But if I just keep doing throw sequences of 10 and only have probabilistic results.

Only in the limit of infinite throws of 10 would one theory go to Probability 1.

~~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.

Note, this a new theory taken as data — so data can be interpreted broadly

The ideal die has exact 6-fold symmetry.

Data. D_3 is exact 6-fold symmetry

$$P(D_3 \mid T_1 \; \cancel{K_2}) = 1$$

$$P(D_3 \mid T_2 \; \cancel{K_2}) = 0$$

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

and $$P(T_1 \mid \cancel{K_3}) = \frac{P(D_3 \mid T_1 \; \cancel{K_2})\, P(T_1 \mid \cancel{K_2})}{P(D_3 \mid T_1 \; \cancel{K_2})\, P(T_1 \mid \cancel{K_2}) + 0} = 1$$

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

What if D_1 acquisitions were throws of 1

$P(D_1 | T_1 C_0) = \frac{1!}{1_1!} \frac{1}{6} = \frac{1}{6}$

(comp. 17045)

$P(D_1 | T_2 C_0) = \frac{1!}{1_1!} \begin{cases} \frac{1}{9} & \text{order} \\ \frac{2}{9} & \text{un} \end{cases}$

So the ideal Bayesian analysis would still be decisive in the limit $\ell \to \infty$.

(comp. 17049)

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

~~But 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 to by principle of indifference and then calculate the Bayes factor

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

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

~~38~~
17052

2025max01

In cosmology they try this combination of data and then that and try to figure out whether there are systematic error 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 ~~error~~ 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.

178

Marginalization

2025 May 01

~~39~~

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 ~~were~~ we are ~~doing~~ keeping general and just need a symbol θ

All of which are continuous variables (i.e. real numbers ...not integers)

except for trivial/toy problems like the die problem ~~P.~~ p. 17045 which perhaps is ~~one~~ almost always ~~for~~

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

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

$$= \frac{P(D_e \mid T_i C_{e-1})}{\langle P(D_e \mid T_j C_{e-1}) \rangle_j}\, P(T_i \mid C_{e-1})$$

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

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

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

2026 jan 24 | 17055

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

But the product $P(D_e|T_i(\theta)C_{e-1})P(\theta|T_i C_{e-1})$ 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 (ie., these are initial conditions not included in parameters) and D_e is exact (ie., error free), then $P(D_e|T_i(\theta)C_{e-1}) = 1$ which it may do for multiple θ and ranges of θ, which doesn't mean $T_i(\theta)$ is true just that it is adequate for the data, 0 otherwise, And there is no case of $P(D_e|T_i(\theta)C_{e-1}) \in (0,1)$

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 <u>cosmology</u> 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 <u>practice</u>, we must expand <u>likelihood</u> in an integral

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

a likelihood probability density

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 to 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.

I calculated $P(D_e | T_i C_{e-1})$ from the data D_e.

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 MCMC, 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 $P(D_2 | T_i C_{2-1})$ 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 errors).

2026jan24 (17059

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.

17060

c) Bayesian Evidence with Marginalization

This is aspect is formally just as on p. 17030, but now with marginalization.

$$OR_{post\,ij} = \frac{P(T_i|C_e)}{P(T_j|C_e)} = \frac{P(D_e|T_i C_{e-1})}{P(D_e|T_j C_{e-1})} \frac{P(T_i|C_{e-1})}{P(T_j|C_{e-1})} = \frac{\int P(D_e|T_i(\theta)\, C_{e-1})\, P(\theta|T_i C_{e-1})\, d\theta}{\int P(D_e|T_j(\theta)\, C_{e-1})\, P(\theta|T_i C_{e-1})\, d\theta} \frac{P(T_i|C_{e-1})}{P(T_j|C_{e-1})}$$

- $OR_{post\,ij}$: Posterior odds ratio
- Bayes K factor of marginal likelihoods
- $OR_{pri\,ij}$: prior odds ratio

In one-step Bayesian analysis $e = 1$ always, but in a single paper the data sets included in $D_e = D_1$ is often varied to see what changes. The data sets usually have varying and uncertain vale.

You use Jeffreys Evidence Table
(see p 17032) or some variation
in order to judge the evidence.

And then, it seems, go on to
an indecisive discussion of what
are the best data sets and
end conclusively/inconclusively.

Alas, competing theories for
the cosmic scale factor $a(t)$
in cosmology are all very close
to each other and the data sets
(e.g., luminosity distances,
angular diameter distances)
though outstanding in size and having
beaten down statistical,
and systematic uncertainties
compared to the past,
are still indecisive in Bayesian analysis

17062

in the overall judgment of the community as far as yours truly can tell as of 2026 May.

It seems some data sets are inconsistent at the level needed for decisiveness suggesting there are uncounted for systematic uncertainties (Heugt et al, 2026 feb 05)

and/or no considered theory is adequate.

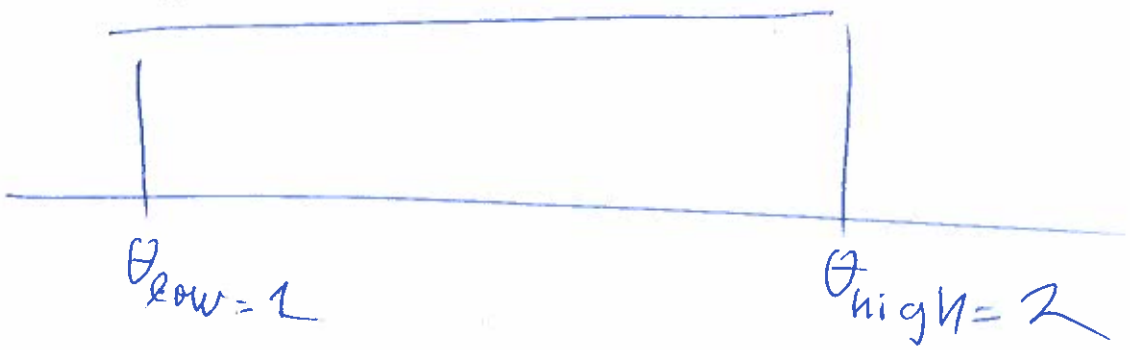
A) Priors: The Prior probability Densities for the Theory parameters

So NOT the initial prior probabilities for the theories $P(T_i | C_{e=0})$ which are usually chosen to be all equal for interesting theories (by the principle of indifference), but $P(\theta | T_i C_{e-1})$ from p. 17054. But choosing the good priors is where physical intuition is needed and that varies widely.

2027 may 01

17063 ~~05~~

The commonest approach is the flat prion



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

You can do more tricky probability densities, ~~distributions~~ but You'd have to have a good reason which could be theoretical (i.e., based on some favored theory)

$\theta_1 \Rightarrow$ 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 | 17064

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

Doing the integral

$$\int P(D_k \mid T_i(\theta) K_{k-1}) P(T_i(\theta) \mid K_{k-1}) d\theta$$

You have to predict for maybe many petabytes of data. Date the likelihood from a multidimensional theory and then do a multidimensional integral

→ It was easy for the die problem (see p. 17045), but it's not easy for e.g. DESI DR2.

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

see p. 17057 discussion

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

Probably, Machine learning is helping.

Also there are ~~simpler~~ 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) ~~Bart~~ The Bayesian Evidence

$$R_{ij} = \frac{P(T_i|K_e)}{P(T_j|K_e)}$$ 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 else [illegible]

~~48~~ | 17066 |

Because of the slop
in the art of priors
and systematic errors,
No one is impressed
by ~~evidence~~ $B_{ij} \cong 10$
maybe not even $B_{ij} = 100$,
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,
unless you have
made bad mistakes,

f) Maximum Likelihood

This is well known technique
for determining the best/likeliest
parameters for a theory $T_i(\theta)$.
If theory is NOT true/adequate,
the parameter values determined may
not be very good or they may have
no meaning relative to the true/adequate theory

2026 jan 24

17067

To find the maximizing parameters

solve

$$\frac{\partial P(D_\ell \mid T_i(\theta) C_{\ell-1})}{\partial \theta} = 0$$

for θ.

But this is a symbolic equation since θ is a set of parameters.

So $\theta \to \{\theta_n\}$ where there N parameters

$$\frac{\partial P(D_\ell \mid T_i(\ldots, \theta_n, \ldots) C_{\ell-1})}{\partial \theta_n} = 0$$

You must solve N simultaneous equations in general

And in nontrivial cases $P(D_\ell \mid T_i(\ldots, \theta_n, \ldots) C_{\ell-1})$ is NOT analytic, and so you forced to use some numerical method (e.g., stochastic gradient ascent, which is analogous to stochastic gradient descent)

17068

You search up gradients, but jump around stochastically so as NOT to get trapped in local Maxima.

For large data sets D_e, this can be computationaly demanding.

And systematic error

or data sets unrepresentative of the true distribution

can lead to bad parameter values

And also if your theory is very inadequate, but has a large number of parameters, you likeliest parameters could be just giving a good fit to noise

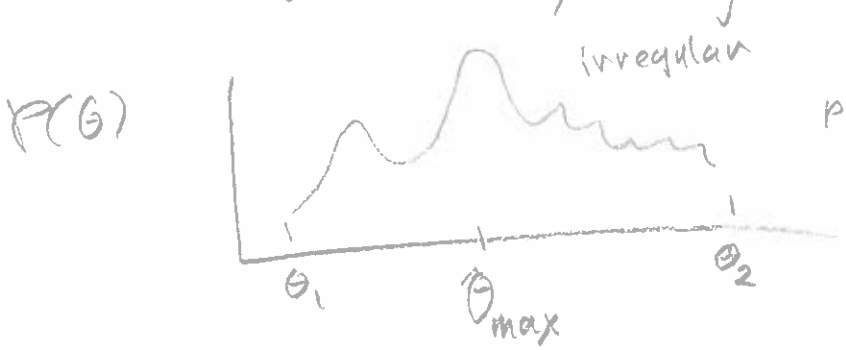
However, if your theory is true/adequate and D_e is sufficiently large, the $\{\theta_n\}$ must asymptotically approach the true values since if you have all possible data the distribution of data must be the true distribution,

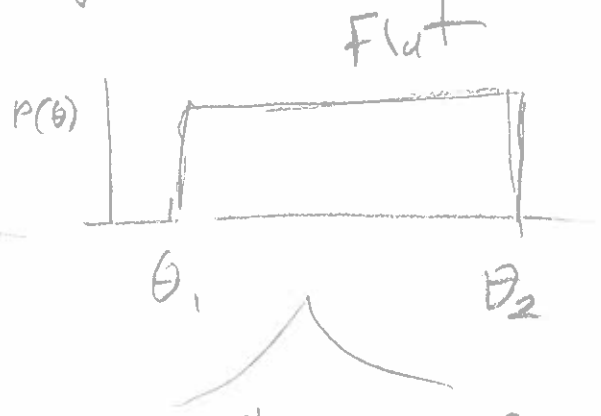
2026 jan 24

17069

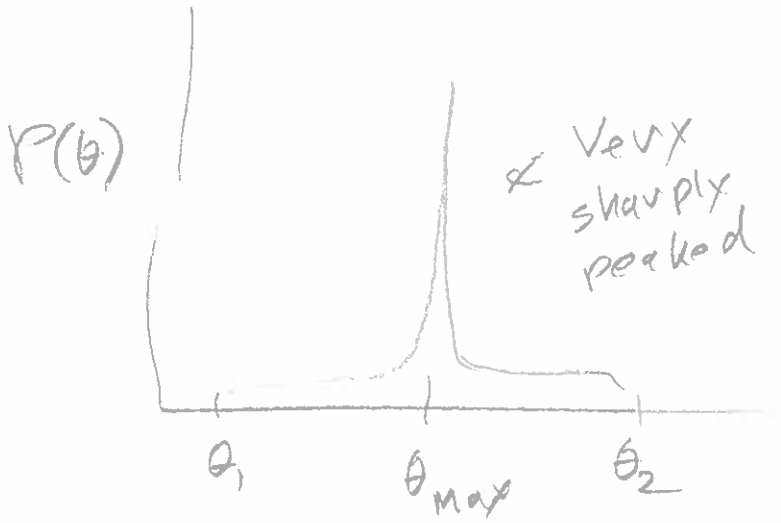
Note, $P(D_2 | T_i(\theta) C_0) = P(\theta)$

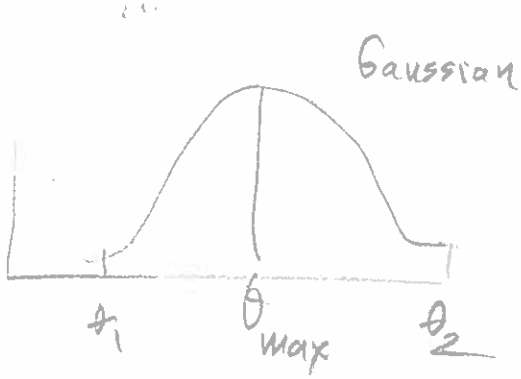
could be anything in general,





The maxima are a continuum





So you may or may not have an easy job finding θ_{max}

and θ_{max} may or may not well approximate θ_{true} if your theory is true/adequate.

Of course, you should follow Rutherford rule and try to get decisive data that allows you to find θ_{max} most easily