Bayesian Analysis (BA)

1) Preamble: BA is path to truth quantified
   a scientific method quantified.

2) Unquantified or Qualitativ Bayesian Analysis
   Priors & Posteriors

3) Bayes Theorem & a bit of History

4) Proof of Bayesian Analysis is in the ideal limit — which can be approach arbitrarily closely
   ~~5)~~ Truth

6) Bayes Odds Ratio & Bayes Factor

7) Multinomial Theorem & Multinomial Probability

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

1) Preamble

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

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

~~Many important theories are like this.~~

at least improvement, The scientific Method Quantified

E.g., Newtonian Physics,

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

One perspective is it an approximate theory

But yours truly (following a herd) 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

Natural selection evolution

unlike Newtonian Physics
~~these~~ can be proven
by mathematical logic
— and so can Bayesian Analysis
aside from pure rigorist quibbling.

2) Qualitative Bayesian Analysis

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

All of life experience gives
vague probabilities about ~~how~~
what thing will happen in ~~many~~
in a given situation
e.g., impossible, virtually impossible,
highly unlikely, unlikely, possibly,
likely, very likely, virtually
certain, certain

4) Those are your prior probabilities
or priors
Then new experience happens
in ~~that~~ situation
and you vaguely update
*our priors
to your posteriors (i.e. posterior probabilities)
(which sounds
awful and often
is)
Your updating is also
qualitative and of variable
accuracy
but works well enough
mostly in everyday life
~~to~~ 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

Ex. Job interviews
especially you are new at the game — or ~~the~~ 'rules' have changed → each interview causes you to update your priors 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 of~~

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

6)

Bayesian analysis itself
is the Sci Method Quantified
(I argue)

3) Bayes Theorem

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

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 proof
~~Bayesian~~ Theorem itself.

I use ABK as joint event or as
shorthand for union = $\cup$ symbol
in math
which is too klutzy for me in this
context.

order has no meaning

Conditional probability

7

$$P(A|K) = \frac{N_{AK}}{N_K}$$

Probability of A given K

Now

$$P(AB|K) = \frac{N_{ABK}}{N_K} = \frac{N_{ABK}}{N_{BK}} \frac{N_{BK}}{N_K}$$

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

a factorization rule which is proven by frequentist approach. But you could assume it as just an axiom — but I think that is pointless.

A frequentist definition of Probability (Trotta p5)

Trotta p.5 argues that need this for probability formalism. But I argue that you need it for probability formalism to have any meaning outside itself.

True N_{AK} and N_K may be vague often, but they must exist in some ideal sense. Idealization is the rule as Galileo argued for understanding.

Clearly

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

∴ Bayes Theorem in symmetric form ← easy to remember

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

Asymmetric forms $P(A|B)P(B) = P(B|A)P(A)$ dropping — or rather just leaving it implicit.

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

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

or supressing K

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

QED.

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

Root of Bayesian analysis

Bayesian analyses was greatly elaboration in 1930s.
(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 from posterior probabilities from theories until one true theory is left.
In toy cases this is easier.
In ~~hard~~ real cases hard, because there is a lot of slop in assigning priors.
No one much used Bayesian analysis before 1990s. Since then it has had growing vogue.

Why was it unused before?

Well, it real power is in dealing with statistical theories (i.e., theories that only give probabilities) and interesting cases in cosmology, epidemiology, social sciences only come with vast data sets and vast computing power to manipulate them. But since 1990s ~~or~~ so 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

speaking in an absolute sense
or having some truth and some falseness definitely in less absolute sense.

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.

7) Proof of Bayesian Analysis in Ideal limit

Say we have ~~background~~ or initial knowledge K_0 about some aspect of reality and a set theories $\{T_i\}$ {I of them} about ~~that~~ aspect. We assume theories discrete for simplicity. Often they form continuums if they are distinguished by free parameter values. In fact such continuums of theories are usually considered one theory with free parameters. But we won't consider that complication now. Later.

[of course, the set is another source part of K_0]

And we have probabilities of being true (to our knowledge)

$$P(T_i \mid K_0)$$ our initial priors

How do you assign $P(T_i|K_0)$

— maybe K_0 gives you an idea

— the barest idea is the Principle of indifference (Kass 1995 p. 776 Not explicit with)

all $P(T_i|K_0)$ are equal.

12

Then you acquire ~~data~~ in ~~D~~ a sequence $D_1, D_2, D_3 \ldots D_\ell$

And your knowledge increase

$$K_1 = K_0 D_1$$

$$K_2 = K_1 D_2 = K_0 D_1 D_2$$

$$\downarrow$$

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

$$\vdots$$

K_∞ = You know everything but don't be disappointed if you don't get there.

Each data acquisition allows you to update priors to posteriors which then become priors for the next data acquisition

Since we are thinking ideally we assume $\{T_i\}$ is complete.

i.e, $P = \sum_i P(T_i | K_0) = 1$

$\hookrightarrow$ if we are wrong we may not find out.

Consider data acquisition ℓ completed and recall Bayes Thm (on p. 7)

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

Prior

Posterior

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

We assume you can calculate the $P(D_\ell \mid T_i K_{\ell-1})$ – the probability of D_ℓ given theory T_i and $K_{\ell-1}$

What is this?

$$P(D_\ell \mid K_{\ell-1})$$

$$= \sum_i P(D_\ell \mid T_i K_{\ell-1})\, P(T_i \mid K_{\ell-1})$$

$$= \sum_i \frac{N_{D_\ell T_i K_{\ell-1}}}{N_{T_i K_{\ell-1}}} \cdot \frac{N_{T_i K_{\ell-1}}}{N_{K_{\ell-1}}}$$

Using ~~the~~ Frequentist definition of probability.
But what does that mean here.

You ~~went~~ did a trial (a complex trial) of reality $N_{K_0} = 2$ times and got 3 ~~theories~~ 3 theories

$N_{K_0} = 3$ since you assume your theories were...

by principle of indifference $\frac{1}{3}$

14) In a vague implicit and hard to calculate sense you get $\frac{N_{TiKe1}}{N_{Ke-1}}$

N's encoded truthily in

and the ratio is easier by far than trying to analyze ~~that~~ the individual N's are — and probably pointless too.

In a sort of multiverse sense N_{Ke-1} universes and N_{TiKe1} theory T_1 is true N_{TiKe1} times

Do I believe this. Sort of but it really takes a more careful

Marginalization

2021nov22

21 (No 15–20 yet)

What if your theories are NOT discrete: i.e. theories $\{T_i\}$ have free parameters collectively symbolized θ?

Bayes Odds ratio

$$\frac{P(T_i(\theta)|K_\ell)}{P(T_j(\theta)|K_\ell)} = \frac{P(D_\ell|T_i(\theta)K_{\ell-1})}{P(D_\ell|T_j(\theta)K_{\ell-1})}\left(\frac{P(T_i(\theta)|K_{\ell-1})}{P(T_j(\theta)|K_{\ell-1})}\right)$$

Bayes K factor

Prior ratio

You don't believe You have all ~~theories~~, even all interesting theories, and so the denominator of Bayesian Analysis is ~~probably~~ usually ~~meaningless and~~ impossible to know — but that exists in principle is vital to Bayesian analysis being ~~ca~~ true theory.

Almost always $\ell = 0$ and these are the initial prior since one only does one explicit Bayesian Analysis iteration. All past knowledge K_0 is implicit earlier iterations.

So you just do Bayes Odds Ratio.

22

But what do you do about free parameters and the priors for a theory with them?

Do you choose the parameters ~~by~~ by maximum likelihood

$\mathcal{L}_{ml} = P(D_e | T_i(\theta_{max}) k_{e-1})$; $\cancel{\mathcal{L}_{max}}$ $\cancel{\mathcal{L}(\theta_{max}, D_e)}$?

~~for data in hand~~

~~If a theory is true, maximum~~

Maximum likelihood gives best choice assuming a theory is true,

$$\lim_{D_e \to \infty} \{\theta_{max}\} = \theta_{true}$$

but not if the theory is not true.

If you assume equal priors for two theories i and j, one with far more free parameters (theory i) then

$$K_{factor} = \frac{P(D_e | T_i(\theta^{large}_{max}) k_{e-1})}{P(D_e | T_j(\theta^{small}_{max}) k_{e-1})}$$

might be spuriously large since the θ_{max}'s are fitted to the D_e.

you could correct by
choosing unequal priors
to put theories on an equal footing
but what is your guidance.
Usually none. In any case θ_{max} may be useless choice of θ if T_i is not true

The path is marginalization. (Implements Occam's razor by effectively eliminating plausibility of parameters of uncertain value and uncertain values)

Expand

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

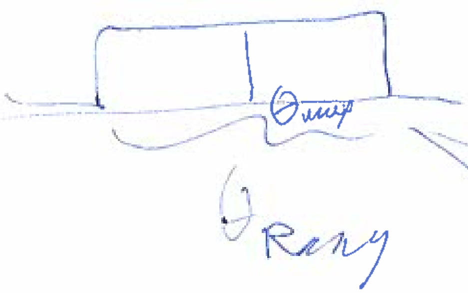
$$= \int P(D_e \mid T_i(\theta) \, K_{e-1}) \, P(\theta) \, d\theta$$

where $P(\theta)$ is a probability density for the priors
Usually, one just guesses

Can do more complicated things but need guidance

$$P(\theta) = \begin{cases} \dfrac{1}{\theta_{range}} & \text{for } \theta \text{ in } \theta_{range} \\ 0 & \text{for } \theta \text{ out of } \theta_{range} \end{cases}$$

The range that you think at all possible



Maximum likelihood parameters are Not favored, since you don't know T_i is true

24)

– as long as you ~~set~~ your range fairly theories with ~~unequal~~ sets of parameters are on an equal footing and you can use equal priors again.

So marginalization implements Occam's razor effectively. (It puts interesting theories on an equal footing.)

↳ eliminating pluralities

Of course for complex theories with many free parameters θ and petabytes or exabytes of data to fit/reproduce ~~Make an unf~~

$$P(D_e \mid T_i k_e)$$

$$= \int P(D_e \mid T_i(\theta) k_e)\, \rho(\theta)\, d\theta$$

can be an immense supercomputer problem.

So it's not exactly choosing formal priors $P(T_i \mid k_0)$ that is the uncertain part — Just set equal for interesting theories. The uncertain part is choosing $\rho(\theta)$ (a sort of prior)

Make an unfair choice
and your Bayes K factor
could be off by orders of magnitude.
→ This why the Jeffreys's scale
sets $\log K = 2$, $(K=100)$
as decisive
and in cosmology $K \approx 10$
is considered completely undecisive
I'd guess.

Bayesian ~~Analysis~~ Analysis, Inference (more standard) 2017 jul 16 [1]

Ideal Procedure for Ch.22 of Cosmology

1) Bayes' Theorem

Bayesian Analysis is a large field of application but theorem is general to all Probability analysis

Proof from frequency analysis (Frequentist analysis)

Is there any other rigorous (non-vague) way?

Yes/No/Maybe

I'd say no at the moment.

What can probability mean if it doesn't correspond to number of events out of a population

A & B are general events, (events = general, e.g., data, theories, beliefs, ...)

K is background event or set of them

All background knowledge ~~count~~ in Bayesian analysis

This is confusing order of these see New p. 34

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

~~Knowledge~~ of all outcomes possible — all data sets. ?, Background knowledge? How to phrase

The theorem is a general simple rule of probability — beyond

$P(AB|K)$

$$P(AB|K) = \frac{N_{ABK}}{N_K} = \frac{N_{ABK}}{N_{BK}}\,\frac{N_{BK}}{N_K}$$

Intersection A and B

$$= P(A|BK)\,P(B|K)$$

Symmetric Form of Bayes Theorem

$P(ABK) = P(A|BK)\,P(BK) = P(A|BK)\,P(B|K)\,P(K)$ Alternative picture → $P(K)$ is usual

$$P(AB|K) = P(A|BK)\,P(B|K) = P(B|AK)\,P(A|K)$$

omiting K as background as usual

$$P(AB) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

Memorable form

Asymmetric Form

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

Omiting K as usual

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

2) It should be obvious how to go from K-omitted form to K form.
But it wasn't initially to me.
Now? Sort of.

Note Bayes' ~~theorem~~ is just a universal rule of probability.

Bayesian analysis as I understand it is ~~the~~ ideal mathematical scientific method which best / most obviously applies to intrinsically probabilistic theories

It's also in qualitative way how we learn ~~[illegible]~~ and all brutes too it seems — including unconsciously

Which most people don't do — But Neural Net machine learn does

2) Bayesian Analysis / Inference

has to be an ideal or how can you trust it

Iterative procedure ideally

Background knowledge $K_0, K_1, \ldots K_\ell$

fresh Data acquired $\cancel{D_0}, D_1, \ldots D_\ell$

a true limit that you can approach

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

The intersection yields new knowledge

ideal limit of all knowledge about a system

a ~~union~~ intersection not a product (of two pieces of data)

It not exhaustive. Never have all $P > 0$. If one true, except true ones but don't know true ones $P > 0$

~~Not~~ Not exhaustive include true ones

Set of theories $\{T_i\}$ (a discrete set for our initial discussion)

Ideally ~~exhaustive~~, which is

Reach limit

B | easily reached when so defined as in student problems

3

$T_i \in K_0$
$P(T_i|K_0) \in K_0$
Both in K_0

Ideally always.

— enough K_0 and there can only be a finite set with non-zero initially probability

Only one theory is true $P(T_i)=1$ if $j \neq i$ $P_j(T_j)=0$

K_0 is not all knowledge — Just all knowledge relative to truth of E_i's

But we don't have that probability about absolute truth ↓ our probabilities are base on our knowledge ~~$P(T_i)$ or probab~~ $P(T_i|K_0)$ is probability our knowledge

Once you have K_0, you know with certainty which theory is true.

in some cases ~~it~~ — so, but sometime it may be small & finite

$P(T_i|K_0)$ initial set of probabilities

Using Bayes Theorem & we always do this Not always

$$l=1 \quad P(T_1|D_1 K_0) = \frac{P(D_1|T_i K_0)\,P(T_i|K_0)}{P(D_1|K_0)}$$

likelihood of getting this data from this theory

likelihood ⇒ in many cases maximum with respect to parameters of the theory

$$l \text{ general } \quad P(T_i|D_l K_{l-1}) = \frac{P(D_l|T_i K_{l-1})\,P(T_i|K_{l-1})}{P(D_l|K_{l-1})}$$

K_l

Now what is $P(D_l|K_{l-1})$ really?

4

It is not $P(D_e)$.

If you have D_e, $P(D_e)=1$.

Always factorize I believe

It's what the universe/reality actually gave you.

Remember ideally ΣT_i is exhaustive

$$P(D_e|K_{e-1}) = \frac{N_{D_e K_{e-1}}}{N_{K_{e-1}}} = \sum_j \frac{N_{D_e T_j K_{e-1}}}{N_{K_{e-1}}}$$

Frequentist understanding

What meaning?

Probability of D given your background knowledge

It's NOT 1.

See

$$= \sum_j \frac{N_{D_e T_j K_{e-1}}}{N_{T_j K_{e-1}}} \frac{N_{T_j K_{e-1}}}{N_{K_{e-1}}}$$

Your understanding in terms of your theory

$P(K_{e-1}$ … $=1$

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

So

Prior

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

Posterior

Which sounds awful we're discussing posteriors

$\sum P(T_i|K_e) = 1$ so every one step normalizes anyway

unless denominator goes to zero.

Weighted average of the $P(D_e|T_j K_{e-1})$

we $\sum_i P(T_i|K_e) \le 1$

Since $P(T_j K_{\ell-1})$ must be normalized, necessary for this to work. Averaging

If $P(D_\ell | T_i K_{\ell-1}) \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} \langle P(D_\ell | T_j K_{\ell-1}) \rangle$

then the Posterior $\left\{ \begin{matrix} \text{increases} \\ \text{stays same} \\ \text{decreases} \end{matrix} \right\}$

(Normalization built in as long as all $P(D_\ell | T_j K_{\ell-1})$ don't go to zero.)

If you choose your new data sets wisely, convergence to

$$P(T_i | K_L) = \begin{cases} 1 & i = I \\ 0 & i \neq I \end{cases}$$

quickly.

(Rutherford Trotta 4)

Choosing wisely on favored theory should help ↓ Rutherford redux ↓ Minimize statistics by doing the right experiment

or data set helps

If not, the convergence can be slow and even terribly misleading. → The "true" theory can have its probability go down.

(see p. 7)

Choose wrong test, bad luck, make blunder

If all data is just statistical with no zero probabilities then $L = \infty$, {But if you are given some non statistical info, L finite.

→ in which case you give up when $P(T_I | K_\ell) = 1 - \varepsilon$

tiny amount

and just say T_I is true unless it fails at some later date

So T_I acts true till your K_ℓ.

It may never do so if ε is super small.

6

Save to p. 17

Now ~~save~~ die problem for example, (see p. 17)

~~$P(T_i | K_0)$~~

If all data statistical, then $L = \infty$ formally.

But say after a million roles, you study symmetry of the die and learn each face is symmetric

$$P(D_e | T_i\, K_{e-1}) = \begin{cases} 1 & \text{for equal probability faces } T_i \\ 0 & \text{otherwise} \end{cases}$$

D_e faces symmetrically

and the iteration stops short.

This is what Rutherford meant about doing wrong experiment

Trotta p. 9

But if your true theory is intrinsically statistical some iteration is necessary.

6a

Example of a case where you may never get ~~to~~ the "truth" in practice.

The iteration stalls till you complete the cycle

Good one

Completely deterministic as to source
but completely random to receiver

Random number generator

Random or completely deterministic

But in some respects, it does NOT matter

In other respects, it does because we want to know truth
(maybe → are we ready for the truth)

It may not matter. But a profound question. Yes profound

Bill Press, Numerical Recipes (1992?)
p. 191 — 199 but seems to have omitted some of the great discussion I recall ✓

for cycle repeat
$Y \cdot 4.3 \times 10^{6001}$ flo
X 10^{17} flops $\times$ $\frac{3 \times 10^{16} s}{Gyr}$

Good one

$\sim \frac{Y}{X} \times 10^{5770}$ Gyr

maybe a quantum computer

Random Number Generators
→ completely deterministic relative to source & will recycle
but random relative receiver

Maybe a quantum computer can solve the cycle?

peta flops, N flops to cycle of Random
$4.3 \times 10^{6001} \cdot N$
$\times 10^{15}$ flops
$\geq 10^{6000}$ Y/X
$\times$ 1 Gyr / 3×10^{16}

Mersenne Twister (MT) pseudo Random
→ $2^{19937} - 1 \triangleq 4.3 \times 10^{6001}$

66

Say you just ~~were~~ were given like string of numbers

After many statistical tests → Maybe a brilliant specially designed test could pull out the determinism, but for most purposes it wouldn't matter

You conclude they are random

in "truth" your wrong → absolutely deterministic

Absolutely wrong effectively right

but for virtually any purpose → Monte Carlo etc.

they work just as if they were.

Philosophically

→ Reality completely deterministic or with intrinsic random element as in QM

Random in time or random out of time in "initial" condition

How can you tell?
Could make any difference to universe evolution or human history?

Maybe Not

3) Bayes Odds Ratio + Bayes Factor

In fact getting the sum of probabilities of theories

Bayesian evidence
Trotta-56

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

(see p. 4)

is hard,

~~But~~ ~~Ethat~~ Average likelihood NOT Max.

Since really one seldom exhaustively knows all theories (~~or all relevant the~~ Well doesn't want to test all anyway)

One often just wants to compare two competing theories ← Relative Probabilities ← Whiscare sovaulen

Practically often only 1 formal iteration Do $\ell = 1$

→ So odds ratio

likelihoods → often maximize to fit parameter ← But in Bayes inference marginalize over parameter

$$\frac{P(T_i|K_\ell)}{P(T_j|K_\ell)} = \frac{P(D_\ell|T_i K_{\ell-1})}{P(D_\ell|T_j K_{\ell-1})}\;\frac{P(T_i|K_{\ell-1})}{P(T_j|K_{\ell-1})}$$

Trotta-57

Posterior Odds ratio (Web) (Wik: Odds Ratio but not explicit)

Kass p.776

Bayes Factor or Bayes ~~k~~ Factor (Wik) or likelihood ratio

Luke Barnes ... in ...

Prior Odds ratio (Web)

often 1 by principle of indifference

3

Jeffrey's ~~log~~ scale ~~scale~~ for k factor

Kass p.776

based on some empirical analysis	< 0	negative for theory i
	$0 - \frac{1}{2}$	barely worth mention
	$\frac{1}{2} - 1$	significant
	$1 - \frac{3}{2}$	strong
	$\frac{3}{2} - 2$	very strong
	$2 \to$ up	decisive

Why? Subject ~~to~~ later revision you do not expect and admit perhaps seeking in a direct sense.

skip to Page 10a!

4) Die Problem (AKA Dice Problem)

Omit in lecture for homework problem

Do multinomial theorem

In general.

I faces labeled $i = 1, \dots I$

$P(T_1)$ Theory 1 $p_i = \frac{1}{I}$

$P(T_2)$ Theory 2 $p_i = \frac{i}{\sum_{i=1}^{I} i} = \frac{i}{\frac{I(I+1)}{2}}$

Priors

Gauss' Proof

$$\begin{array}{lll} 1 & I & = I+1 \\ 2 & I-1 & = I+1 \\ \vdots & & \\ I & 1 & = I+1 \end{array}$$

$P(T_3)$ Theory 3 $p_i = \frac{2^{\text{mod}(i,2)}}{\frac{3I}{2}}$ modulo function (WiK) assume I even,

so evens twice as likely as odds

simpler so ~~I 1+1+2~~ ~~= 3 3 tota~~

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

Say N throws $= N = \sum_i n_i$

with outcomes $n_1, n_2, \dots n_I$

Don't

$$P(\{n_i\}) = \frac{N!}{n_1! \, n_2! \dots n_I!} \prod_i p_i^{n_i}$$

multinomial theorem used to get

Throws 1, 1, 2, 4, 1

only one ordering occurs

$C \, n_1! \dots n_I! = N!$ among identical throws

Those that look identical only occur once identical permutations actually occur

Combinations Wik: Combinatity [illegible]

10

$$1 = (P_1 + P_2 + \cdots + P_I)^N$$

Total probability of outcomes repeating N times

$$= \sum_{\text{all sets}} \frac{N!}{n_1! \ldots n_I!} \prod_{i=1}^{I} P_i^{n_i}$$

Wik Multinomial Theorem

$$= N! \prod_{i=1}^{I} \frac{P_i^{n_i}}{n_i!}$$

~~all combinations turn up once only because none different by symmetry~~

All branchings occur,

- all combinations ~~or "arrangements"~~
- $N\{$ = number of combinations ~~"arrangements"~~

P_1 P_2 P_1 P_2 $\frac{L_2}{L}$ $\frac{L_1}{L}$ $\frac{L_3}{L}$

reorder among identical counts

order is counted, among identical — Permutation: order matters

ordering among identical is NOT counted → reordering among identical not part of path.

only select from 1st bound not ordered

Combination — order does not matter — How to understand that order when exactly chosen in order

only order counted among identical

1^{st} — Prob of whole event I — Nth

- each path is (e.g. $P_1 P_2 P_I P_{I-1} P_4 \ldots P_3$) is the probability of that path. (prod of each item)

~~arrangement~~ combination

- each path occurs once only → combination all look different, → Permutations of identicals don't.

by symmetry

- (combinations) Paths of equal probability (the distinct subset) ~~are all the same~~ combination (Not counting accidental equalities)

Number of combinations $\{n_i\}$ same value

$$C = \frac{N!}{n_1! \ldots n_I!} \leftarrow C(n_1, n_2, \ldots, n_I, N)$$

distinct subset

4) Multinomial Theorem & Multinomial Probability Distribution

– A digression to let me set up an example application of Bayesian analysis

Say you have a set of J events: 1, 2, ... J
with probabilities $P_1, P_2, \ldots P_J$

And you then took a (sequence) sample of N ~~trials~~ events

and got, e.g., P_2 P_{J-3} P_4 P_7 ... P_{J-1} ... P_J = Product is the prob this

Order event count → 2, J-3, 4, 7, ..., J-1, J
1 2 3 4 ... N-1 N

~~How many distinct combinations are there~~

distinct permutation

Permutation Period!

statistics jargon

combination: collection of items without order distinction

permutation: collection of items with order distinction

In our case distinct permutation have a set probability already

distinct permutations: Those you can tell apart by appearance.

For N objects of J kinds
~~For our trials~~

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

Permutation of N ~~objects~~ objects

Number of ~~combos~~ distinct permutations ~~corresponding~~ of combination $\{n_i\}$ ← multinomial coefficient

Number of permutation among identical objects.

10b

So $C(\{n_i\}) = \dfrac{N!}{\prod n_i!}$ — i.e. number of distinct permutations of contribution $\{n_i\}$

Now consider our trials.

All distinct permutations ~~tur~~ can turn up ~~and~~ ~~are distinct overall events~~, and are individual events

But any ~~indist one~~ "set" of indistinct ones are just one overall event.

...1...2...

...2...1...

Two overall events

You draw from the 1 bin, then the 2 bin and then the reverse

Two ways of doing this

...1...1...

...1...1...

Just one overall event.

Here you draw from the 1-bin then 1-bin

Just one way of doing it!

Consider

	1	2	3	4
event	1	1	2	6
prob	P_1 *	P_1 *	P_2 *	P_6

⟹ the probability of this event

Imaginary interchange of 1's doesn't double the probability. (Selection with replacement in stats jargon)

Probability of the ~~collection~~ combination $\{n_i\}$s

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

every distinct permutation of $\{n_i\}$ has this product probability

$\therefore$ Multinomiale ~~D~~ Distribution

$$\mathcal{P}(\{n_i\}) = \frac{N!}{\prod_i n_i!} \prod_i p_i^{n_i}$$

Prove

$$= N! \prod_i \left(\frac{p_i^{n_i}}{n_i!}\right)$$

Now

total probability conserved

$$\sum_i \mathcal{P}(\{n_i\}) \stackrel{?}{=} 1$$

It should be

Proof The pobability of some event happening is 1.

Well $1 = \sum_i p_i$ on first event of N sequence

$p_i \sum_j p_j$

$1 = \left(\sum p_i\right)^2$ on second event of N sequence

then sum on i

$\vdots$

$1 = \left(\sum_i p_i\right)^N$ on N^{th} event of N sequence

$p_i \left(\sum_j p_j\right)^2$ then sum for three events and so on

10d

Multinomial theorem

$1 =$ expansion is isomorphic to our counting of the $C(\{n_i\})$ $= \sum_i P(\{n_i\})$

with ~~with~~ one-to-one correspondence with $p_1, p_2, p_3 \ldots p_3$ replacing $1, 2, 3, \ldots 3$

I'm not totally convinced of the cogency, but the result is true QED

~~Special case~~ Binomial Thm : 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 $q = 1-p$

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

the n choose k formula

Just called binomial coefficient (Wiki)

Moments of the binomial distribution.

A old trick

Define function

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

$$M_\ell = \left(x\frac{d}{dx}\right)^\ell f\Big|_{x=p} = \sum_{k=0}^{n} k^\ell \binom{n}{k} x^k q^{n-k}\Big|_{x=p} = \left(x\frac{d}{dx}\right)^\ell (x+q)^n\Big|_{x=p}$$

10e

$l=0 \quad M_0 = 1$

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

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

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

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

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

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

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

$\cancel{= p^3 n(n-1)} = pn - p^2 n$

$= pn(1-p)$ BeV-53 correct

1 of

$r_i = \frac{1}{(\ldots)}$ $\sum r_i = 1$ as it should

Now sampled N events and got $\{n_i\}$

Theory 1

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

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

Theory 3

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

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

There for I events I even $P_i \propto 1$ i odd $P_i \propto 2$ Σ Total $= \frac{I}{2}\cdot 1 + \frac{I}{2}\cdot 2 = \frac{3}{2}I$

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

Theory 2

$$P(\{n_i\}) = \frac{N!}{n_1!\ldots n_I!}\,\frac{1}{\left(\frac{3}{2}I\right)^{N}}\prod_{i=1}^{I}\left[2-\mathrm{mod}(i,2)\right]^{n_i}$$

Very easy to post dict $P(D|T)$ not like complex system with huge data

5) Die Problem for Zeus II

Numerical problem

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

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

six items

$P_1 + P_2 + \ldots + P_6$

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

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

Prior odds $P(T_1)/P(T_2) = 1$

Principle of indifference (Barnes p.6)

So 10 Throws $(N=10)$ complex enough, but ~~not~~ decisive?

I really did this

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

math 2

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

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

12)

$K = 3.604$

Probability given your knowledge — Before even, No assumption

So Prior odds = 1

Bayes factor $\left\{ \begin{array}{l} k = 3.604 \\ \text{Posterior odds} = 3.604 \end{array} \right.$

If ~~they~~ two theories really are exhaustive, one can calculate their probabilities

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

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

$P(1) \Rightarrow 1$

$P(2) \Rightarrow 0$

But my die is not perfect and it can be loaded. Face 4 opens!! {Stopped it}

$P_1(1) = 0.7828..$, $P_2(2) = .21799..$ 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

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

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

So in math 2

we get $k = 3.604$... again

because same number of even & odds

But pattern odds sup.?

$= k \times \left(\frac{P(1, k_1)}{P(1 \mid k_1)}\right) = k^2 \simeq 13$

But favours theory 1 twice in a row

$P_2(1k_2) = .9285$... $P_2(2k_2) = .0714 \,8$...

So do we keep advancing to truth by more & more ~~trials~~? trials?

But no! Instead of rolling die, we investigate its symmetry

Rutherford Rule

$D_3 \Rightarrow$ ~~6~~-fold symmetry

$P(D_3 \mid 1k_2) = 1, \quad P(D_3 \mid 2k_2) = 0$

$$P(1 \mid K_3) = \frac{\overset{1}{P(D_3 \mid 1k_2)}\, P(1 \mid k_2)}{P(1 \mid k_2) \cdot \underset{1}{P(D_3 \mid 1k_2)} + P(2 \mid k_2)\, \underset{0}{P(D_3 \mid 2k_2)}}$$

$P(1, k_2)$ & $P(2, k_2)$ are now irrelevant

$= 1$

keep ~~4~~ but factor in other order!

What if you can't or don't know how or it's just needed a definitive single experiment.

$$P(2 \mid k_3) = \frac{P(D_3 \mid 2k_2) * P(2 \mid K_2)}{\ldots}$$

$= 0$

So Rutherford was right!! Trotted.

"If you need statistics, you did the wrong experiment!!"

(4) But if the system is intrinsically probabilistic (Die is)

and you can not find ~~its~~ distribution directly (as you can for ideal Die)

then Bayesian analysis makes sense.

~~Perhaps~~ — Cosmology

— social science

— epidemiology

Marginalization proof

5) Marginalization

allowed parameter range

< density

$dP_x = \frac{dx}{g}$

$\int \langle D_e | T_i K_{e-1} \rangle \langle T_i | K_{e-1} \rangle dx$

$T(x)$

integrate over all allowed range

- some predict data well some low

Say your theory has a free parameter θ

it's a continuum infinity of theories.

But best to regard as one theory with free parameters

Set of θ.

Maybe be common choice:

Trotta's choice for set of free param

If you think your theory

$P(D_e | T_i K_{e-1})$ r.i.
$P(D|\theta M)$

is right fit it to | 15

the data using maximum likelihood to determine free parameters. (All data relevant to theory → all

Very θ to maximize

(global one or local if that are possible

We'll ~~do this~~ expand on this is later.

Local maxima & global maxima

Several theories all with free parameters

as ~~K~~ → all
$\Theta \rightarrow \Theta_{exact}$

if your theory is exactly right, all data should be fit

to you can be fooled by either or both

the best fit one may just because it has lots of free parameters

Chi-square well degrees of freedom correction

3eV-314 = reduced Chi-Square

But also if your theory has enough free parameters and you are fooled – wonder off path. What if Two different theories fit absolutely all Data → something equivalent seen by some

Even wrong theory? maximum theory

Need Occam's razor

compare Non-best fits of theories

compare Best fits

Bayesian always us to integrate over free parameters over good range the prior range

And put ~~complex~~ many & few parameter theories on a level playing field.

Trotta - 57 says a level more parameters eualized – maybe

and rate theories ~~independent~~ independent ~~of~~ best ~~of~~ parameter fit

just in actual practice rather than in principle

16) — rate on best fit
how prior range of parameter.

→ Some a nuisance parameters

Can't be measure maybe

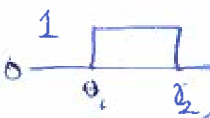
→ not of interest (Wik)
at least at moment
and deciding on the theory

So now

$$P(T_i \mid K_{\ell-1})$$

$$= \int P(T_i(\theta_i) \mid K_{\ell-1}) P(\theta_i) d\theta_i$$

Often Top hat prior



But Truth is 30

θ_i parameter of theory$_i$

weight function

} a probability distribution itself

of $P(\theta) = \begin{cases} 1 & \text{where possible} \\ 0 & \text{otherwise} \end{cases}$

where you know the parameter can't be

But
Replace p.4 formula.

LHS

$$\frac{P(D_\ell \mid T_i(\theta_i) K_{\ell-1}) P(T_i(\theta_i) \mid K_{\ell-1}) P(\theta_i) d\theta_i}{\sum_j \int P(D_\ell \mid T_j(\theta_j) K_{\ell-1}) P(T_j(\theta_j) \mid K_{\ell-1}) P(\theta_j) * d\theta_j}$$

sum over all ~~the others~~

Marginalize over all parameter sets

Flat prior Not Best!

$P(T_i(\theta_i) \mid K_\ell) d\theta_i =$

RHS

$P(T_i \mid K_\ell)$

17

If you have vast sets of data

Cosmological peta bytes

1 pb = 10^{19} bytes

LHC in 5 pb per year

couple process 24 pb per day in 2009

— Integration can be huge even for our modern standard just doing K factors

— Are shortcuts of varying utility of which I so far know little

(1) Bayesian Information Crit (WIC (WIK

eg Schwarz crit

= BIC

Finding likelihood $P(D | T_i(\theta), K)$

for the problem simple but petabytes produced from a complex theory

↓

I intuit that here you need Markov Chain Monte Carlo MCMC (Trotta-35 and a lot skill

derived from Bayesian approx

18

derive from information theory

Akaike Information Crit. (Wik

AIC

relative

quality of model

Not absolute

smaller better

from information theory

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

$k\uparrow$ $AIC\uparrow$

number of estimated parameters

max likelihood of model

$L\uparrow, AIC\downarrow$

for number of parameters the same

$\left[\ln\left(\frac{L_{max}}{L_{max}, AIC \text{ in min}}\right)\right]$

$\simeq$ k or ~~odds~~ posterior odds ratios

By someone, proof

Best of set ~~of AIC~~

$[AIC_{~~min~~} - AIC]/2$

$= ~~[2\ln L_{max} - 2\ln L]~~/2$

$= 2\ln(L/~~L_{max}~~)/2 \simeq$

like k or Posterior odds

has 1, others lower

Probably better than comparing ~~Chi~~ reduced Chi-Square

~~Pro~~ > 1 which become indicative of $p < 0.05$ or more

BIC NOT

somehow related

Similar to ~~Bayesian~~

some

different weighting

~~They~~ BIC & AIC must work well in some ~~ideal~~ limit, & so useful

but can't be universally use or people