AN EDUCATIONAL NOTE ON BAYESIAN ANALYSIS

David J. Jeffery[1]

ABSTRACT

Bayesian analysis (or Bayesian inference if you prefer) is a quantified theory of learning and the scientific method. In this note, we prove that Bayesian analysis is a true theory in an ideal limit that can be approached sufficiently closely for Bayesian analysis to be useful. In fact, we posit without proof that any useful theory should be true in an ideal limit that can be approached sufficiently closely to be, as said, useful. Some ancillary results for Bayesian analysis are also presented.

Subject headings: supernovae: cosmology: theory — cosmological parameters — dark energy

1. INTRODUCTION

The schematic scientific method is a cycle of theorizing and observation (which often includes experimentation or computer simulation). In the cycle, wrong theories and wrong observations are ideally eliminated and if you have access to all possible relevant theories and observations ideally you eventually arrive at true about whatever aspect of reality you are studying.

However, a complication is that often theories cannot be ruled in or out by a new set of observations. Rather their probabilities goes up or down depending new set. Very often these probabilities are judged ver qualitative: e.g., "after data X, theory Y is now very likely/unlikely." In fact, the procedure of updating from prior probabilities (priors) to posterior probabilities (posteriors) has always gone on in everyday life. For example, say you are going to a job interview. You have prior expectations about things will go. After the interview update your expectations based on your experience. This is all done usually qualitatively although sometimes the posteriors have gone to flat zero: e.g., you expect to get the job and you don't.

[1]Department of Physics & Astronomy and Nevada Center for Astrophysics (NCfA), University of Las Vegas, Nevada, 4505 S. Maryland Parkway Las Vegas, Nevada 89154, U.S.A.

Bayesian analysis quantitifies the cycle of updating priors to posteriors, and thus quantifies the scientific method. The ideal Bayesian analysis with access to all possible relevant theories and observations can be proven to lead to truth: i.e., the true theory of the aspect of reality you are studying. We give the proof in § 4. Note the word "ideal." Practicable Bayesian analyses cannot guarantee arrival at truth. However, practicable Bayesian analyses can approach the ideal Bayesian analysis closely enough that Bayesian analysis has become highly useful in certain fields of research in which the theories themselves predict probabilities: e.g., cosmology, sociology, psychology, epidemiology, and economics.

Why if Bayesian analysis is useful was not used much until the 1990s (depending on how one counts things). It's useful when you have large data sets and enormous computer power to analyze them. Such data sets and computer power have grown together since the 1950s gradually making Bayesian analysis more and more practicable in nontrivial applications.

Now the basic element of Bayesian analysis is Bayes' theorem which was discover by Thomas Bayes in 1763 and independently by Pierre-Simon Laplace in 1774 (Wikipedia: Bayes' theorem). It is an extremely simple theorem and is simply proven which we do in § 2. Bayesian analysis itself is application of Bayes' theorem to the analysis of data. The very basic ideas of Bayesian analysis are very simple as we will show in § 4. The developments beyond the basic ideas and the applications of Bayesian analysis are immense fields that we do not go into in this note. Much of the pioneering development of Bayesian analysis was done by Harold Jeffreys (1891–1989) (Wikipedia: Harold Jeffreys; Jeffreys 1961)—no relation though he manages to look like my father anyway.

The contents of this note are as following. In § 2, we prove Bayes' theorem as aforesaid. § 4 § 5 § 6 § 7 § 8 § 9

2. BAYES' THEOREM

Here we prove Bayes' theorem which is the basic theorem of Bayesian analysis. Consider three general events A, B, and K. The event K is not necessary to the proof, but it is useful to include it for proving Bayesian analysis itself. When we get to that proof, K will stand for background knowledge. The joint probability for A, B, and K is $P(ABK)$, where here we use the "product" to stand for intersection of events. We factorize the probability $P(ABK)$:

$$P(ABK) = P(A|BK)P(BK) = P(A|BK)P(B|K)P(K) \ , \tag{1}$$

where $P(U|V)$ means is the conditional probability of U given V. Similarly,

$$P(ABK) = P(B|AK)P(AK) = P(B|AK)P(A|K)P(K) \ . \tag{2}$$

Equating the last two results gives Bayes' theorem in symmetrical and unsymmetrical forms:

$$P(A|BK)P(B|K) = P(B|AK)P(A|K) \tag{3}$$

$$P(A|BK) = \frac{P(B|AK)P(A|K)}{P(B|K)} \qquad P(B|AK) = \frac{P(A|BK)P(B|K)}{P(A|K)} \ . \tag{4}$$

If we suppress the K as unnecessary, we get usual expressions for Bayes' theorem in symmetrical and unsymmetrical forms:

$$P(A|B)P(B) = P(B|A)P(A) \tag{5}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad P(B|A) = \frac{P(A|B)P(B)}{P(A)} \ . \tag{6}$$

3. MEANING OF PROBABILITIES

I believe that the meaning of probability is always the ratio of some events/outcomes to some number of trials for the events/outcomes in the limit of infinite trials. This is the frequentist definition of probability (e.g., Trotta 2017, p. 5). A Bayesian perspective on probability is that probability can also measure just degree of belief, and so can deal with unique events and so is more general than the frequentist definition (e.g., Trotta 2017, p. 5). However, this Bayesian perspective seems lacking meaning to me although you can develop much probability formalism without the frequentist definition.

Rather than argue my position in general, I will just take consider a case that occurs in the next section, § 4, and which I think generalizes easily to complete generality. Say initially we have a set of theories $\{T_i\}$ about some aspect of reality based on background knowledge K_0 relevant to the set of theories. Only one theory at most can be true logically: the others are false. How can we assign probabilities to the theories (i.e., $P(T_i|K_0)$) with only one trial reality? Well we can imagine an infinite number of possible realities in which one obtains K_0, but where the other aspects of the realities cover all possibilies and can be quite different. Those imagined realities imply that the probabilities $P(T_i|K_0)$ do exist. Can we know them? Well if theory T_j is ruled out or in by K_0, then we know already $P(T_j|K_0) = 0$ or $P(T_j|K_0) = 1$ even with only having one reality that leads to K_0. In the ruled-in case, we even know $P(T_{i\neq j}|K_0) = 0$. What if no theory is ruled in, but some are possible by K_0. (Hereafter, I only refer to the set of possible theories $\{T_i\}$.) There may be some way we can know exact probabilities of $\{T_i\}$ in special cases: e.g., we know them by definition in the toy case in § 6. But in most realistic cases, we will not know exactly the nonzero $P(T_i|K_0)$. But I emphasize they do exist and we can estimate them by some piece of information included

in K_0. The estimation may be based how similar cases work have worked out in the past. If nothing else and we deem all of set $\{T_i\}$ plausible, there is the principle of indifference: just assign them all equal probability. (e.g., Wikipedia: Principle of indifference).

Now one might argue for the Bayesian perspective of degree of belief by saying the estimation of the $P(T_i|K_0)$ is the degree of belief. However, I think one still needs the argument of the last paragraph to give meaning to "degree of belief," and so one is basing "degree of belief" on a frequentist definition of probability ultimately.

The frequentist definition of probability implies that joint probabilities can always be factorized. For proof:

$$P(AB) = \lim_{N\to\infty} \frac{N_{AB}}{N} = \lim_{N\to\infty} \frac{N_{AB}}{N_B}\frac{N_B}{N} = P(A|B)P(B) \ ,$$

where A and B are general events, N_{AB} is the number of joint AB events, N_B is the number of B events, and N is the number of trials. We used the factorization of probabilities in the proof of Bayes' theorem in § 2.

Note that Bayes' theorem itself demonstrates that there are other ways of expanding probabilities besides factorization.

4. PROOF OF BAYESIAN ANALYSIS

In this section, we prove that Bayesian analysis in the ideal limit (i.e., that ideal Bayesian analysis) leads to a true theory a general phenomenon X. Note that ideal Bayesian analysis is an iterative procedure.

For general iteration step ℓ, we have knowledge K_ℓ about X which includes a set of possible theories $\{T_i\}$. We also know the theory probabilities $P(T_i|K_\ell)$ for step ℓ. The initial knowledge is K_0 and the initial possible theory probabilities are $P(T_i|K_0)$. The initial probabilities in most realistic cases will be estimates and not exactly known values, but those are all we need for the ideal Bayesian analysis. See the discussion of estimated theory probabilities in § 3. The initial probabilities are our initial prior probabilities or initial priors.

Note that knowledge K_ℓ may not include all possible theories consistent with K_ℓ and one of the not included ones may even be the true theory. This all right. The procedure corrects for this deficiency. However, in order expand in the $P(T_i|K_\ell)$ (which we must do: see below), we need to treat the known set $\{T_i\}$ as being be exhaustive for K_ℓ even if it is actually not. Again the procedure corrects for this deficiency. Because the known set $\{T_i\}$ is being treated as exhaustive, the probabilities $P(T_i|K_\ell)$ can be normalized and, also for

expansion, we require that we have normalized the probabilities: i.e., we have

$$1 = \sum_i P(T_i|K_\ell) \ . \tag{7}$$

The iteration procedure consists in aquiring new data sets D_ℓ: i.e., $D_1, D_2, \ldots D_L$. Every iteration, we update our knowledge: $K_\ell = K_{\ell-1}D_\ell$, where again the "product" stands for intersection of events. Note that one might usually think of K_ℓ as a union of $K_{\ell-1}$ and D_ℓ. However, if we think of $K_{\ell-1}$ and D_ℓ as events, then having them both is an intersection. We define D_L to be when we have enough knowledge that $P(T_i|K_L) = 1$ for the true theory T_i and all the other probabilities are zero. Note endpoint L could be finite or infinite.

Clearly, we have to introduce new theories suggested by the D_ℓ if the initial set $\{T_i\}$ did not include the true theory. New theories might be introduced even if the initial set did include the true theory since since we do not know the true theory before the endpoint. We must estimate new probabilities for newly introduced theories and renormalize the set of probabilities.

For iteration ℓ, we have posterior probabilities

$$P(T_i|K_\ell) = P(T_i|K_{\ell-1}D_\ell) = \frac{P(D_\ell|T_iK_{\ell-1})P(T_i|K_{\ell-1})}{P(D_\ell|K_{\ell-1})} \ , \tag{8}$$

where we have used Bayes' theorem equation (4) changing the variables names, *mutatis mutandis*. As we from equation (8) that the prior for theory T_i at iteration ℓ gets updated to the posterior by the use of Bayes' theorem.

The quantity $P(D_\ell|T_iK_{\ell-1})$ is the likelihood (e.g., Wikipedia: Likelihood function), but it cannot be varied by varying free parameters since there are none for our theories T_i. Note that the $K_{\ell-1}$ in $P(D_\ell|T_iK_{\ell-1})$ cannot be suppressed as irrelevant since observational/experimental setup giving D_ℓ goes into the calculation of $P(D_\ell|T_iK_{\ell-1})$ which we assume that we can do in the ideal Bayesian analysis.

Now note that we have the data set D_ℓ in hand, and so we have the absolute probability $P(D_\ell) = 1$. So what is $P(D_\ell|K_{\ell-1})$ actually? It's the probability of obtaining D_ℓ given our knowledge $K_{\ell-1}$. Since our set of theories is exhaustive, we can expand $P(D_\ell|K_{\ell-1})$ in those theories:

$$P(D_\ell|K_{\ell-1}) = \sum_i P(D_\ell|T_iK_{\ell-1})P(T_i|K_{\ell-1}) = \langle P(D_\ell|T_iK_{\ell-1})\rangle \tag{9}$$

where the second expression just recognizes that the sum gives us the mean likelihood for iteration ℓ given the priors are normalized. Note $P(T_iK_{\ell-1}) = P(T_i|K_{\ell-1})P(K_{\ell-1}) = P(T_i|K_{\ell-1})$ since $P(K_{\ell-1}) = 1$ since we do, in fact, have $K_{\ell-1}$ in hand.

We can now rewrite equation (8) for the posterior probabilities as

$$P(T_i|K_\ell) = \frac{P(D_\ell|T_iK_{\ell-1})}{\langle P(D_\ell|T_iK_{\ell-1})\rangle} P(T_i|K_{\ell-1}) \ . \tag{10}$$

So if $P(D_\ell|T_iK_{\ell-1})$ greater/lesser than the mean likelihood, theory T_i gains/loses in probability in the ℓth iteration step. Now the ideal Bayesian analysis proceed iteration step by iteration step using equation (10). Some theories gain probability, some lose, some go to zero probability and are discarded, and sometimes new theories are introduced. There may steps where all theories have zero probability, but we can keep collecting new data D_ℓ and inventing new theories.

How do we know when the endpoint has been reached? Let's enumerate some cases:

1. In some cases, we might know by logical necessity. These might often be toy cases (e.g., see § 6) or, at least, contrived cases. But there are important cases there are true theories by logical necessity. A prime example example is the theory of evolution by natural selection. As a mechanism, it is can be proven to work on the computer. The DNA hardware in biota is there to implement it in nature, and so it must hold in nature—as a vast amount of observational data shows too.

2. In some cases, by pure exhaustion: you knowledge K_L is exhaustive about the phenomenon X. If we have one nonzero probability theory left at exhaustion, it must be true. If we have multiple nonzero probability theories left at exhaustion, they must all be true and somehow equivalent. What if you have no theory at exhaustion? Then K_L itself constitutes a theory though not a very elegant one.

3. We may become exhausted when we reach step ℓ and have a vast knowledge K_ℓ for X and a theory T_i completely adequate for X. We can then define ℓ to be step L. In this case, we can say we have an adequate theory for X relative to K_L. If we have multiple theories all adequate for K_L, then they are equivalent in some sense at least for K_L.

   An important special case of this case is when theory T_i is a special case of a much broader theory which is adequate for a vast realm of phenomena beyond X. All that vast realm then verifies theory T_i for its much smaller realm. For example, Newtonian physics is adequate for all motion in the classical limit (i.e., much larger than atoms, much slower than the vacuum light speed, much weaker gravity than black holes) and it follows as limiting form of quantum mechanics (so it is thought), special relativity, and general relativity which theories are adequate for much broader realms. Newtonian physics is thus a highly adequate theory and one can just say it is true in the classical limit since wherever it fails is not sufficiently close to the classical limit.

4. Maybe the endpoint is out of reach in practice. Consider the random number generator the Mersenne Twister (Wikipedia: Mersenne Twister). It has a repeat period of $\sim 10^{6622}$ and passes many standard tests for randomness. Say we can only do those standard tests and do not know the source of the random numbers. Given an exascale computer (Wikipedia: Exascale computing), how long would it take to exhaust the period and know all the numbers were determined completely deterministically and not completely randomly? Of order

$$t \approx \frac{10^{6622}}{10^{18}\,\mathrm{s}^{-1}} = 10^{6596}\ \text{Julian years}\ . \tag{11}$$

   In this Mersenne-Twister case, we would spend long ages without being able to decide definitely or even probably whether the numbers were generated randomly or completely deterministically. In practice, Bayesian analysis fails.

   Actually, the idea of a nearly perfect deterministic random number generator raises an interesting philosophical point. There are two theories about reality: 1) it is a combination of deterministic and intrinsic quantum mechanical randomness; 2) it is completely deterministic with quantum mechanics being actually completely deterministic as in some unconventional quantum mechanics theories. But if the events can be completely deterministic as to source, but completely random as to receiver, how could one ever tell which theory about reality is true? And even if one could tell, would it matter very much to reality. Reality determined by either theory might be much the same, and so the distinction between them may not be as important as one might think.

Does one ever do ideal Bayesian analysis? Probably only for toy cases as we do in § 6. In fact, overwhelmingly most scientific advance has never used Bayes' theorem nor thought of itself as a Bayesian analysis. However, in the scientific method the updating from priors to posteriors goes on all the time qualitatively without being classified as qualitative Bayesian anslysis which is what it is. How could this informal qualitative Bayesian analysis work. Well, there is an aphorism attributed to Ernest Rutherford (1871–1937): "If you need statistics, you did the wrong experiment" (e.g., Trotta 2017, p. 4). The essense of this aphorism is that decisive experiments/observations rule theories out (i.e., their posteriors go to zero), and so accelerate the iteration to a true theory without needing to calculate nonzero probabilities. Certainly in formal Bayesian analysis, it makes sense to accelerate the iteration by choosing to obtain the most decisive set of data D_ℓ you can find.

Given the last paragraph, why does one need (formal) Bayesian analysis at all? When dealing with theories that give only statistical predictions and differ in their predictions not

vastly, Bayesian analysis is the best tool for proceeding to the best theory. Such theories turn up in, as mentioned in the § 1 in, e.g., cosmology, sociology, psychology, epidemiology, and economics. Bayesian analysis though known to some degree for a long time (with much pioneering development already done by 1961: Jeffreys 1961), only became relatively important in research the later 20th century when vast computing power and powerful numerical techniques became available to make it practicable.

In fact in practice, Bayesian analysis is often only one iteration and just uses ratios of posteriors to rank theories two at time. Say we have theories T_i and T_j and we obtain data D_1. Then using equation (10), we can write

$$R_{\mathrm{po}(ij)} = \frac{P(T_i|K_1)}{P(T_j|K_1)} = \frac{P(D_1|T_iK_0)}{P(D_1|T_jK_0)}\frac{P(T_i|K_0)}{P(T_j|K_0)} = B_{ij}R_{\mathrm{pr}(ij)} \ , \tag{12}$$

where $R_{\mathrm{pr}(ij)}$ is called the prior odds ratio, $R_{\mathrm{po}(ij)}$ is called the posterior odds ratio, and B_{ij} is called the Bayes factor (e.g., Kass & Raftery 1995, p. 776). If the priors are estimated by the principle of indifference (e.g., Wikipedia: Principle of indifference), the prior odds ratio is 1 and the posterior odds ratio equals Bayes factor: i.e., $R_{\mathrm{po}(ij)} = B_{ij}$.

In fact, the usual practice seems to be to a not-ideal Bayesian analysis with an informal iteration where the prior odds ratio set to 1 at every step. The theories in this procedure are given more study depending on how they are ranked by the posterior odds ratios: if highly, then a lot; if lowly, then little or none; if in between then in between study. After more study, the more favored theories from the first analysis are ranked again, and so on. Note the iterations are often carried out by different researchers and are not thought of as iteration.

Why do such not-ideal Bayesian analyses rather than something much more formal? Very probably because there are usually only very crude estimates of prior probabilities and, in most cases, crude estimated ranges for the free parameters (see § 7). This means that a more formal Bayesian analysis will give false precision, and so is pointless. Proof of this statement is that the Bayesian analyses of a phenomenon by different researchers often comes to very different posterior odds ratios.

A consequence of the lack of precision of Bayesian analysis in practice is that Bayes factors of order a few are not considered significant in most cases in ruling out theories. Kass & Raftery (1995, p. 777) has given a table rating Bayes factor evidence against a theory based on general expectations as they see them. We give their table in Table 1 below with the reasonable generalization to Bayes factor B_{ij} or posterior odds ratio $R_{\mathrm{po}(ij)}$ evidence against theory T_j. We call this general ratio R_{ij}. As Table 1 shows, a theory T_j would have to have a very high ratio R_{ij} against to overcome the uncertainty in estimated priors before one could even provisionally discard it from consideration.

Table 1. Bayes factor or posterior odds ratio evidence R_{ij} against theory T_j

$2\ln(R_{ij})$	R_{ij}	Evidence against j
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

5. THE MULTINOMIAL DISTRIBUTION

In order to give our toy case of Bayesian analysis in § 6, we need to use the multinomial distribution AKA the multinomial probability distribution (Wikipedia: Multinomial distribution). So we introduce it here.

Say we have a set of I possible events:

$$1, 2, 3, \ldots, I \quad \text{with individual probabilities} \quad P_1, P_2, P_3, \ldots, P_I \quad \text{with} \quad \sum_i P_i = 1 \tag{13}$$

We now take a sample of N events and obtain

$$i, j, k, \ell, \ldots, m \qquad \text{with joint probability} \qquad P_i P_j P_k \ldots P_m \ . \tag{14}$$

The distinct permutations of the events are different samples which have their own joint probability equal to the one shown. The indistinct permutations of events is just the same sample. To count up the probability for the set of disinction permutations note

$$N! = C(\{n_i\})\Pi_i n_i! \ , \tag{15}$$

where $N!$ is the number of permutation of events, the number of identical events of type i is n_i, $\{n_i\}$ is the set of identical events in the sample, and $C(\{n_i\})$ is the number of distinct permutations for set of identical events $\{n_i\}$. Solving for $C(\{n_i\})$ gives

$$C(\{n_i\}) = \frac{N!}{\Pi_i n_i!} \tag{16}$$

Thus, the collective probability for the set of disinction permutations is

$$P(\{n_i\}) = C(\{n_i\})\Pi_i P_i^{n_i} = \frac{N!}{\Pi_i n_i!}\Pi_i P_i^{n_i} = N!\Pi_i \left(\frac{P_i^{n_i}}{n_i!}\right) \tag{17}$$

which is, in fact, the multinomial distribution.

The sum of the $P(\{n_i\})$ for samples of size N should be 1. What else could it be? But if we need a proof to satisfy paranoia, behold:

$$1 = \sum_i P_i = \left(\sum_i P_i\right)^N = (P_1 + P_2 + \ldots + P_I)^N = \sum_i P(\{n_i\}) \tag{18}$$

where expanding the third member of the equation via the fourth member to get the fifth member is isomorphic to construction of the $P(\{n_i\})$ in the paragraph above when you think about it really hard.

As pointless digression, let's consider the special case of the binomial distribution where there are only two events which we assign probabilities p and q with $p + q = 1$, of course. Starting from equation (18), we see that

$$1 = (p+q)^N = \sum_{k=0}^{N} \binom{N}{k} p^k q^{N-k} \ , \tag{19}$$

where the binomial distribution is

$$P(N,k) = \binom{N}{k} p^k q^{N-k} \tag{20}$$

and binomial coefficients follow from equation (16):

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} \ . \tag{21}$$

The moments of the binomial distribution follow from the old trick of replacing p by variable x in the normalization equation (i.e., eq. (19), differentiating ℓ times by $\ln(x)$, and evaluating the ℓth derivative with $x = p$ times for the ℓth moment. Behold:

$$\begin{aligned}
f(x) &= \sum_{k=0}^{N} \binom{N}{k} x^k q^{N-k} = (x+q)^N & (22)\\
M_\ell &= \left[\left(\frac{d}{d\ln(x)} \right)^\ell \sum_{k=0}^{N} \binom{N}{k} e^{k\ln(x)} q^{N-k} \right] \Bigg|_{x=p} = \sum_{k=0}^{N} k^\ell \binom{N}{k} p^k q^{N-k} \\
&= \left[\left(\frac{d}{d\ln(x)} \right)^\ell (e^{\ln(x)} + q)^N \right] \Bigg|_{x=p} & (23)\\
M_0 &= 1 & (24)\\
M_1 &= \left[N(e^{\ln(x)} + q)^{N-1} e^{\ln(x)} \right] \Big|_{x=p} = Np & (25)\\
M_2 &= \left[N(N-1)(e^{\ln(x)} + q)^{N-1} e^{2\ln(x)} + N(e^{\ln(x)} + q)^{N-1} e^{\ln(x)} \right] \Big|_{x=p} \\
&= N(N-1)p^2 + Np & (26)\\
\sigma^2 &= Np(1-p) & (27)
\end{aligned}$$

which results are confirmed by, e.g., Bevington (1969, p. 53)

6. A TOY CASE OF BAYESIAN ANALYSIS

For a toy case of Bayesian analysis using the procedure and results of § 4, say we have a standard die (singular of dice): i.e., a cube with i dots on a side: i ranges from 1 to 6 and

the total number of possible events on a trial (i.e., a throw of the die) is $I = 6$. We are given two a priori theories of the probability distribution that determines the up-side of throw:

$$\text{Theory } T_1: \qquad P_i = \frac{1}{6} \qquad \text{for all } i \tag{28}$$

$$\text{Theory } T_2: \qquad P_i = \frac{2 - \text{mod}(i,2)}{9} = \begin{cases} \dfrac{1}{9} & \text{for } i \text{ odd;} \\ \dfrac{2}{9} & \text{for } i \text{ even.} \end{cases} \tag{29}$$

Note function $\text{mod}(a, n)$ is remainder of a divided by n, where the remainder is not divided by n: the function is vocalized a modulo (or mod) n. The basic knowledge about the die and the two theories are out initial knowledge K_0 Using the principle of indifference, we assign prior proabilities

$$P(T_1|K_0) = P(T_2|K_0) = \frac{1}{2} \tag{30}$$

which gives a prior odds ratio of 1.

We start the Bayesian analysis iteration by accumlating data set 1 which consists of 10 throws of the die with results as follows:

throw count	1	2	3	4	5	6	7	8	9	10
dots on up-side	1	5	1	1	2	5	2	6	3	4 .

The likelihoods for the data set for the two theories are computed using the multinomial distribtution equation (17) introduced in § 5 with $N = 10$, $n_1 = 3$, $n_2 = 2$, $n_3 = 1$, $n_4 = 1$, $n_5 = 2$, $n_6 = 1$:

$$P(D_1|T_1K_0) = \frac{10!}{(3!)(2!)(1!)(1!)(2!)(1!)} \left(\frac{1}{6}\right)^{10} = 2.50057\ldots \times 10^{-3} \tag{31}$$

$$P(D_2|T_1K_0) = \frac{10!}{(3!)(2!)(1!)(1!)(2!)(1!)} \left(\frac{1}{9}\right)^{10} \times 1^6 \times 2^4 = 6.9238\ldots \times 10^{-4} \tag{32}$$

which yield the posterior odds ratio and Bayes factor as shown in the following equation

$$\frac{P(T_1|K_1)}{P(T_2|K_1)} = \frac{P(D_1|T_1K_0)}{P(D_1|T_2K_0)} \times 1 = Bij = 3.604\ldots \ . \tag{33}$$

Taking Table 1 at face value, the Bayes factor R_{ij} is positive evidence against theory T_2, but is far from being strong evidence against.

We will now do a second iteration of the Bayesian analysis in which the posterior probabilities of the first iteration

$$P(T_1|K_1) = 0.7828006\ldots \qquad \text{and} \qquad P(T_2|K_1) = 0.217199\ldots \tag{34}$$

are the prior probabilities of the second. We obtain data set 2 which also consists of 10 throws of the die with results as follows:

throw count	1	2	3	4	5	6	7	8	9	10
dots on up-side	6	1	3	6	6	2	1	5	3	5 .

Again the likelihoods for the data set for the two theories are computed using the multinomial distribtution equation (17) introduced in § 5 with $N = 10$, $n_1 = 2$, $n_2 = 1$, $n_3 = 2$, $n_4 = 0$, $n_5 = 2$, $n_6 = 3$:

$$P(D_1|T_1K_1) = \frac{10!}{(2!)(1!)(2!)(0!)(2!)(3!)}\left(\frac{1}{6}\right)^{10} = 1.250\ldots\times 10^{-3} \quad (35)$$

$$P(D_2|T_1K_1) = \frac{10!}{(2!)(1!)(2!)(0!)(2!)(3!)}\left(\frac{1}{9}\right)^{10}\times 1^6\times 2^4 = 3.469\ldots\times 10^{-4} \quad (36)$$

which yield the posterior odds ratio and Bayes factor as shown in the following equation

$$\frac{P(T_1|K_2)}{P(T_2|K_2)} = \frac{P(D_1|T_1K_1)}{P(D_1|T_2K_1)}\frac{P(T_1|K_1)}{P(T_2|K_1)} = Bij\frac{P(T_1|K_1)}{P(T_2|K_1)} = (3.604\ldots)\times(3.604\ldots) = 12.989\ldots\ . \quad (37)$$

It seems perfectly reasonable

7. FREE-PARAMETER THEORIES, MARGINALIZATION, AND OCCAM'S RAZOR

8. AIC AND BIC

9. CONCLUSIONS

Conclusions are in the abstract and *Introduction* (i.e., § 1).

Support for this work has been provided the Department of Physics & Astronomy and the Nevada Center for Astrophysics (NCfA) of the University of Nevada, Las Vegas.

A. Appendix A

REFERENCES

Bevington, P. R. 1969, Data Reduction and Error Analysis for the Physical Sciences (New York: McGraw-Hill Book Company)

Jeffreys, H. 1961, Theory of Probability (Oxford: Clarendon Press)

Kass, R.E., & Raftery, A.E. 1995, Journal of the American Statistical Society, 90, 773

Trotta R. 2017, arXiv:1701.01467