A silhouette of Bayes' theorem

S. Adarsh[1]*

Abstract
This article briefly describes the history and essential workings of Bayes' theorem. The article does not delve into mathematical subtleties or the metaphysical aspects of the theorem. From his experience, the author feels that understanding the basic philosophy of any mathematical tool upfront is often a compelling experience and would spur one to delve deeper into the subtleties later. To further understand the application of Bayes' theorem, a numerical example and its MATLAB code are also included.

Keywords
Bayesian probability; frequentist probability; probabilistic updating;

[1] *Ph.D. Candidate, Dept. of Civil Engg., Indian Institute of Technology Kanpur, India*
***Corresponding author**: adarshss@iitk.ac.in

Contents

1. A tale of two probabilities

Two schools of thought exist for probability: 1) frequentist probability and 2) Bayesian probability. A description of these two types of probabilities are given below.

1.1 Frequentist probability

In frequentist probability, the probability of an event, A, is defined as the relative frequency of the event over a large number of trials. This definition is formally expressed as:

$$p(A) = \lim_{N_t \to \infty} \frac{N_A}{N_t} \tag{1}$$

Here, N_A is the number of occurrences of the event, A, in N_t number of trials. The caveat of such a definition is that the probability can only be defined for an event for which trials can be performed. For example, this kind of interpretation cannot to be directly applied to find the probability of a candidate winning an election, because conducting the elections over a number of trials is not feasible. Still, this is the most common interpretation of probability.

1.2 Bayesian probability

1.2.1 A new language in probability acquires grammar

Bayes' theorem was conceived in its incipient form by Reverend Thomas Bayes (see Fig. 1a) in the 1700s. For some reason, Bayes did not publish his finding during his life. Luckily, his work [1] was published posthumously by his friend Richard Price and is available to the public gratis. Later, Pierre-Simon Laplace (see Fig. 1b) propounded the modern form of the Bayes' theorem in 1774 and showed how it could be applied to science. Some believe that the contributions of Laplace is just as important as that of Bayes. Throughout history, this theorem has been used for solving many complicated problems. One noted application was its use by Alan Turing, the father of computer science, to crack the Nazi's infamous Enigma code. So, Bayes' theorem also played a part in defeating Hitler. Bayes's theorem was also used by Jerome Cornfield in the field of medicine to establish an indubitable correlation between lung cancer and smoking. Despite these successful applications, the world of statistics had shunned this theorem for many decades.

This is primarily due to the British polymath Ronald Fisher vehemently campaigning against it. He probably knew that some of his significant contributions, like the maximum likelihood estimator, were just corollaries of Bayes' theorem. Fisher might have been an excellent mathematician, but he might not have been a pleasant character to be around with. It was only towards the end of the 20$^{\text{th}}$ century that Bayes' theorem became widely accepted. It must be mentioned that the books (not for tyros) by R.T. Cox [2] and E. T. Jaynes [3] played a pivotal role in the revival of Bayes' theorem. Apart from them, Arthur Bailey, I.J. Good, Dennis Lindley, and Jimmie Savage also gave impetus to the resurrection of Baye's theorem. Mcgrayne[4] provides a good account of the history and development of the Bayes' theorem.

(a) **(b)**

Figure 1. (a) Thomas Bayes and (b) Pierre-Simon Laplace , source: www.wikipedia.org.

1.2.2 The coveted equation

In Bayesian philosophy, the probability of an event is always conditioned on some available information/data. This probability is interpreted as the quantified value of belief in that event with the data provided. The idea here is to assume an initial belief for an event and then keep on updating this initial belief by using real-life data to get improved beliefs. Frankly, this interpretation of probability seems more intuitive and natural to the author than the frequentist interpretation. Bayes' theorem for the probability of an event A conditioned on information, B, is given as:

$$p(A/B) = \frac{p(A)p(B/A)}{p(B)} \tag{2}$$

where $p(A)$ is the probability of A before acquiring the information B; $p(B/A)$ is the probability of B given A and $p(B)$ is the probability of B. Unlike frequentist probability, no trials are required to get $p(A/B)$. Once the information of B is obtained, the mathematical model that correlates A and B can be used to get $p(A/B)$. In other words, $p(A)$ is updated to $p(A/B)$ using the information B. Even though Bayes' theorem is just a statement about conditional probability, its implications are quite profound. It gives a structured method to deduce the unknown event A provided we have information about B and the relationship between B and A

Now suppose we have N_c candidates and we want to find the probability of the i^{th} candidate winning in an election given the data B collected from small samples within the population. The probability is given as:

$$p(A_i/B) = \frac{p(A_i)p(B/A_i)}{\sum_{i=1}^{N_c} p(A_i)p(B/A_i)} \tag{3}$$

In the above equation, $p(A_i)$ is the probability of the i^{th} candidate winning without any information from data which is usually a subjective guess. The probability $p(B/A_i)$ is usually calculated with the mathematical model that correlates the data with the event that the i^{th} candidate wins. The theorem that equates $p(B)$ to the summation in the denominator of Equation 3 is known as the theorem of total probability.

A more common and practical version of Bayes' theorem is the one that is based on probability density functions (PDFs), rather than scalar probabilities, expressed as:

$$p(\boldsymbol{\theta}/\mathscr{D},\mathscr{M}) = \frac{p(\boldsymbol{\theta}/\mathscr{M})p(\mathscr{D}/\boldsymbol{\theta},\mathscr{M})}{p(\mathscr{D}/\mathscr{M})} \tag{4}$$

By using the above equation, our objective is to find the PDF of parameter $\boldsymbol{\theta}$ of a model $\mathscr{M}$ based on the information carried in the measured data $\mathscr{D}$. A model is a mathematical form that relates $\boldsymbol{\theta}$ and $\mathscr{D}$. In Equation 4, $p(\boldsymbol{\theta}/\mathscr{D},\mathscr{M})$ is known as the posterior PDF of $\boldsymbol{\theta}$; $p(\boldsymbol{\theta}/\mathscr{M})$ is the prior PDF of $\boldsymbol{\theta}$; $p(\mathscr{D}/\boldsymbol{\theta},\mathscr{M})$ is the likelihood function; $p(\mathscr{D}/\mathscr{M})$ is the probability of the data given the model, calculated as: evaluation of $p(\mathscr{D}/\mathscr{M}_i)$ as:

$$p(\mathscr{D}/\mathscr{M}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}/\mathscr{M})p(\mathscr{D}/\boldsymbol{\theta},\mathscr{M})d\boldsymbol{\theta} \tag{5}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^{N_\theta}$, and N_θ is the number of parameters in model $\mathscr{M}$. If we are dealing with only one model, the evaluation of $p(\mathscr{D}/\mathscr{M})$ is not necessary, i.e., we only has to deal with the numerator of Equation 4. The PDF $p(\boldsymbol{\theta}/\mathscr{M})$, in the numerator, incorporates subjective information, while the function $p(\mathscr{D}/\boldsymbol{\theta},\mathscr{M})$ incorporates the objective information which includes the mathematical model and the real-life data. So, a significant feature of this theorem is that we can fuse subjective and objective information to estimate unknown parameters of a model.

Let us suppose we have identified $p(\boldsymbol{\theta}/\mathscr{D},\mathscr{M})$ for a model where $\boldsymbol{\theta} = \{\theta_1, \theta_2\}^T$, and this identified distribution is plotted in Fig. 2. The projection of the probability distribution onto the $\theta_1 - \theta_2$ plane is illustrated in Fig. 3. It can be observed that Bayesian identification, unlike classical identification, gives multiple possibilities of $\boldsymbol{\theta}$ for a given $\mathscr{D}$, and the plausibility of each of these possibilities is given by the corresponding value of $p(\boldsymbol{\theta}/\mathscr{D},\mathscr{M})$. As an example, three points of $\boldsymbol{\theta}$ are chosen at random, as indicated by the cross marks in Fig 3, and their corresponding values of $p(\boldsymbol{\theta}/\mathscr{D},\mathscr{M})$ are enumerated in Table 1. These multiple possibilities arise due to epistemic uncertainty, and the explanation of this uncertainty is beyond this article's scope.

So far, we have only considered one model that relates $\boldsymbol{\theta}$ and $\mathscr{D}$. A notable consequence of Bayes' theorem is that a quixotic model that can thoroughly explain data does not exist. This means that multiple models that can explain the data exist, and the relative plausibility of a model $\mathscr{M}_i$ among N_m models is calculated as:

$$p(\mathscr{M}_i/\mathscr{D}) = \frac{p(\mathscr{M}_i)p(\mathscr{D}/\mathscr{M}_i)}{\sum_{i=1}^{N_m} p(\mathscr{M}_i)p(\mathscr{D}/\mathscr{M}_i)} \tag{6}$$

The class of problems described by Equation 4 is known as Bayesian parametric identification, while the class of problems described by Equation 6 is known as Bayesian model identification. Bayesian model identification supersedes Bayesian parametric identification and warrants the evaluation of $p(\mathscr{D}/\mathscr{M}_i)$ for each model through Equation 5.

Table 1. The three points chosen and the values of the posterior distributions

Point No.	θ_1	θ_2	$p(\boldsymbol{\theta}/\mathscr{D},\mathscr{M})$
1	1.8	1.8	0.12
2	2	1.2	0.10
3	3	1.1	0.04

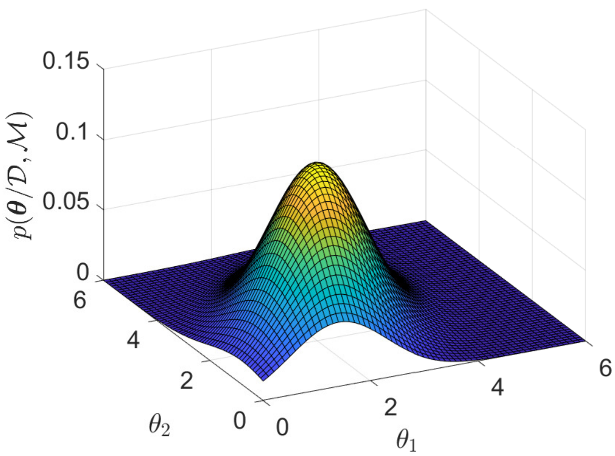

Figure 2. The plot of the identified posterior distribution, $p(\boldsymbol{\theta}/\mathscr{D},\mathscr{M})$

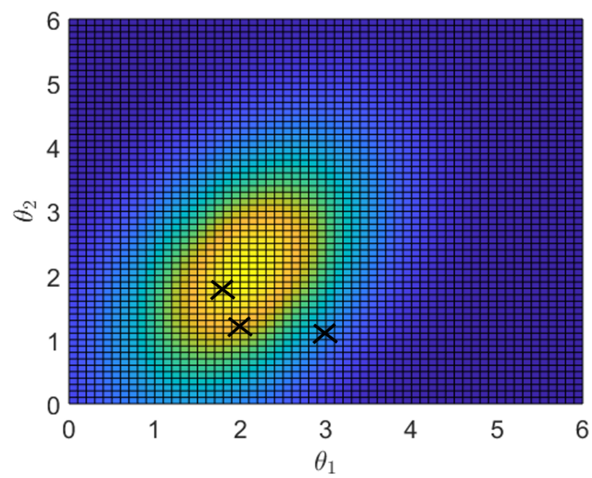

Figure 3. The projection of the probability distribution onto the $\theta_1 - \theta_2$ plane, and the three points of $\boldsymbol{\theta}$ that are chosen.

2. Parallels of Bayes' theorem with the human mind

Bayes' theorem works in a manner quite similar to the human mind's thinking process. Imagine yourself as the boss of a multinational conglomerate, and you want to interview and hire a new employee from a list of candidates. Firstly, you did some thorough background checks and perusal of their curriculum vitae. So, before the interview itself, you have some initial beliefs about each candidate. These initial beliefs are equivalent to the prior probability distributions in Baye's theorem. Now, during the interview, you keep updating the initial beliefs based on the performance of the candidates. Of course, this updating depends on a slew of factors like the questions asked, the ability of the candidate to manipulate you, and the time taken for the interview. This updating of beliefs based on observed real-life data is achieved through the likelihood distributions in Bayes' theorem. After updating your initial beliefs through the interviews, you select a candidate based on your updated beliefs. These updated beliefs are comparable to the posterior probability distributions in Bayes' theorem. Now, this process can be continued, i.e., your beliefs can be repeatedly updated down the line based on the candidate's performance in your company. If the candidate continues to perform well consistently, you can give him a raise, or on the contrary, if his performance continues to be abysmal, you can fire him. However, there is one subtle difference, how the human mind quantifies beliefs is not lucid, while Bayes' theorem provides a systematic and cogent way to quantify beliefs through probabilities.

3. Example: Estimation of the bias of a coin by Bayes' theorem

In this section, the estimation of the bias of a coin by Bayes' theorem is illustrated and is based on a simulation run on MATLAB [5]. Bias essentially quantifies the number of particular outcomes one gets in 100 trials. For example, if a coin has a bias of 0.3 for heads, one can expect to get around 30 heads for 100 flips. So, naturally, an unbiased coin will have a bias of 0.5 for both heads and tails. The corresponding MATLAB code for the simulation and plotting of the ensuing results are given in Appendix A and can also be downloaded from `https://github.com/adarshss-github/Bayesian_Coin_Toss`. After fixing the coin's bias for getting heads, the flips of the coin are simulated, and the data acquired from the simulation are used to update a prior PDF of the bias. So, essentially, the objective is to find the updated PDF, representative of the bias, after each simulated coin flip. It should be noted that frequentist probability is more congenial for a problem like flipping a coin. However, Bayesian probability is used here just for illustrative purposes.

The prior PDF of bias, before the first flip, is assumed to be a triangular distribution that peaks at an arbitrary value of bias, b, and is denoted as:

$$p(P_H/\mathscr{M}) = \begin{cases} (2/b)\times P_H & \text{if } P_H \in [0,b] \\ (2/(1-b))\times(1-P_H) & \text{if } P_H \in (b,1] \end{cases} \tag{7}$$

While the likelihood function after N coin flips and N_H occurrences of heads is:

$$p(\mathscr{D}/P_H,\mathscr{M}) = \frac{N!}{N_H!(N-N_H)!}P_H^{N_H}(1-P_H)^{(N-N_H)} \tag{8}$$

Moreover, Equation 8 is nothing but the well-known binomial distribution. The likelihood function quantifies the likelihood of getting N_H heads in N flips, given that the bias for heads is P_H. It should be noted that the data set, $\mathscr{D}$, provides information about N and N_H after each trial. So, the posterior PDF for bias is given by:

$$p(P_H/\mathscr{D},\mathscr{M}) = \frac{p(P_H/\mathscr{M})p(\mathscr{D}/P_H,\mathscr{M})}{p(\mathscr{D}/\mathscr{M})} \tag{9}$$

After the first coin flip, $p(P_H/\mathscr{M})$, is updated to $p(P_H/\mathscr{D},\mathscr{M})$. The denominator of Equation 9 can be calculated by the condition given in Equation 5. $p(P_H/\mathscr{M})$ evaluated from the first flip is used in the prior PDF for the second flip. Using the posterior PDF from the previous flip as the prior PDF in the ensuing flip can be continued for an arbitrary number of flips.

For the illustration, the coin is first allocated a bias of 0.1 for heads, and 1000 flips are simulated. A triangular prior PDF that peaks at a bias of 0.9 is chosen. As discussed before, after each flip, the data is used to update the posterior PDF, and this updated posterior PDF is used as the prior PDF in the subsequent flip. Figure 4 shows the prior PDF before six particular flips. It can be seen from Fig. 4 that the prior PDF is gradually updated such that the prior PDF before Flip 1000 is concentrated over the bias of 0.1. Further, it can also be observed that the uncertainty in the updated PDF reduces as the number of flips increases and more information becomes available. The reduction in uncertainty manifests as the reduction in the spread of the PDFs and an increase in the heights of the PDFs. Had one started with a prior PDF with a peak nearer to 0.1, lesser information or fewer flips would have been required to achieve the same results as that for a peak near 0.9.

The Bayesian and frequentist estimation bias are plotted in Fig. 5. In Fig. 5, the Bayesian bias is calculated as the maxima of the posterior PDF, and it can be seen that the results of both types of estimation match as the number of flips increases. Videos showing the updating results from Figs. 4 and 5 can be accessed at `https://youtu.be/rg7XFSpTguQ`, and `https://youtu.be/f7qw3MP8uPg`, respectively.

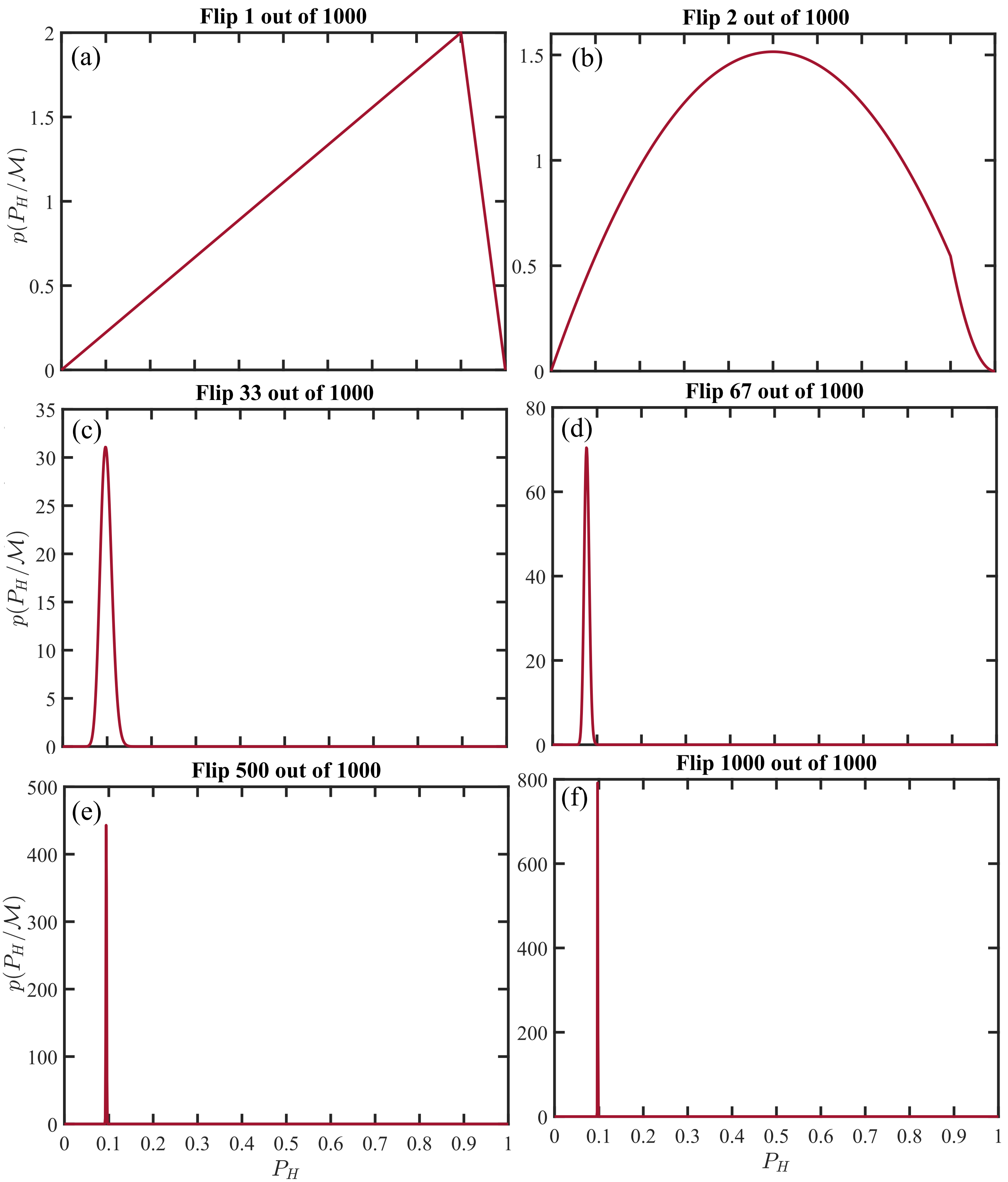

Figure 4. The prior PDF before particular flips. (a) Flip 1, (b) Flip 2, (c) Flip 33, (d) Flip 67, (e) Flip 500, and (f) Flip 1000.

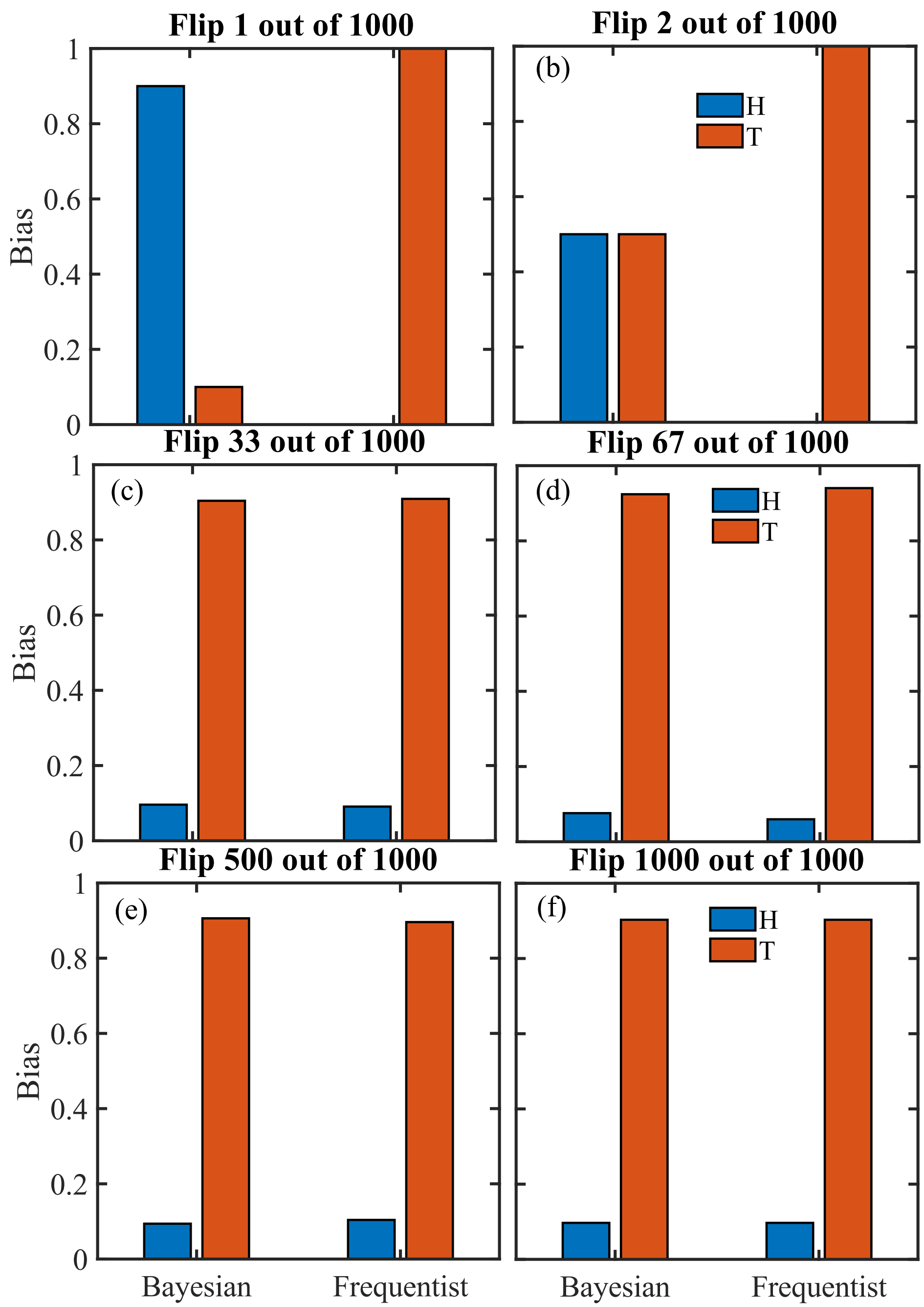

Figure 5. Comparison of bias by Bayesian and frequentist methods after particular flips. (a) Flip 1, (b) Flip 2, (c) Flip 33, (d) Flip 67, (e) Flip 500, and (f) Flip 1000.

References

[1] BAYES. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3-4):296–315, 1958.

[2] Richard T Cox. The algebra of probable inference. *American Journal of Physics*, 31(1):66–67, 1963.

[3] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[4] Sharon Bertsch McGrayne. *The theory that would not die*. Yale University Press, 2011.

[5] The MathWorks Inc. Matlab version: 9.13.0 (r2022b), 2022.

A. MATLAB code for estimating and plotting the bias of a coin

```

%By Adarsh S, Ph.D. Candidate IIT Kanpur
%Email: adarshss@gmail.com
%
%Script description:
%-----------------------
%This MATLAB script demonstrates Bayesian updating for evaluating the bias
%of a coin. A triangular prior PDF is assumed and the peak of the prior can
%be set to an arbirary value. The bias is set upfront and a particular number
% of coin flip are simulated. The prior PDFs are successively updated by
%using the results from the flip to get the posteriror PDF of the bias, through
%Bayes theorem.
%
%
%Note:
%------
%The updated priors after each flip is stored in prior_PH_MAST
%The outcomes of the trials are stored in Flips (1 for heads and 0 for tails )
% After running Section 1, one may proceed to evaluate Sections 2 and 3 to
% plot the results

clear all
clc
%%
%SECTION 1:
%-----------
%This section simulates the coin flips and updates the prior PDFs by Bayes
%theorem

%1) Set Bias of coin (say Head)
%------------------------------
%------------------------------
bias_Coin = 0.1 ;
%2) Set number of coin flips
%------------------------------
%------------------------------
coin_Flips = 1000 ;
%3) Set peak of traingular prior PDF
%------------------------------
%------------------------------
peak_Prior = 0.9 ;
%------------------------------
%------------------------------

```

```
PH = 0:0.001:1 ;
prior_PH = zeros(1,length(PH)) ;

for i = 1:1:length(PH)

prior_PH(i) = prior_P1_TRI_bia(PH(i),peak_Prior) ; %Peak of prior

end

trapz(PH,prior_PH) ;

Tria = coin_Flips ; % Number of flips
prior_PH_MAST = zeros(1,length(PH),Tria) ;
post_PH = zeros(1,length(PH),Tria) ;
lik_PH = zeros(1,length(PH),Tria) ;
prior_PH_MAST(1,:,1) = prior_PH ;
NH = zeros(1,Tria) ;

bias_PH = bias_Coin ; %Bias of coin
Flips = double(rand(Tria,1) < bias_PH) ;
Flips = Flips' ;

for i = 1:1:Tria

NH(i) = length(find(Flips(1:i)==1)) ;

for j = 1:1:length(PH)

lik_PH(1,j,i) = like_PH(i,NH(i),PH(j)) ;
i
j

end


post_PH(1,:,i) =  (prior_PH_MAST(1,:,i).*lik_PH(1,:,i))/trapz(PH,
   prior_PH_MAST(1,:,i).*lik_PH(1,:,i))  ;

if i<Tria

prior_PH_MAST(1,:,i+1) = post_PH(1,:,i) ;

end

end


%%
%SECTION 2 (plots):
```

```
%-----------
%This section plots the updated prior PDFs after each flip

for i = 1:1:Tria
figure(1)
plot(PH, prior_PH_MAST(1,:,i),'LineWidth',2,'Color',[0.6350 0.0780 0.1840])
title(sprintf('Flip %d out of %d', i, Tria ),'FontSize',15,'FontName','Times New Roman')
xlabel('$P_H$','interpreter','latex','FontSize',15,'FontName','Times New Roman')
ylabel('$p(P_H)$','interpreter','latex','FontSize',15,'FontName','Times New Roman')
set(gca,'Xtick',0:0.1:1,'fontname','times')
ax = gca;
ax.XRuler.TickLabelInterpreter = 'tex';
ax.FontSize = 15;
ax.TickLength = [0.02,0] ;
ax.LineWidth = 2;
drawnow


end

%%
%SECTION 3 (plots):
%----------
%This section plots the comparison of probability estimated through
%Bayesian and frequentist approaches

for i  = 1:1:Tria
figure(1)

[~,I] = max(prior_PH_MAST(1,:,i)) ;
[~,I] = max(post_PH(1,:,i)) ;
PHB = PH(I) ;
PTB = 1 - PHB ;

PHF = NH(i)/i ;
PTF = 1 - PHF ;

bar([PHB PTB; PHF PTF],'LineWidth',1.5) ;
xticklabels({'Bayesian','Frequentist'})
title(sprintf('Flip %d out of %d', i, Tria ))
ylabel('Bias','FontSize',15,'FontName','Times New Roman')
legend('H','T')
legend('boxoff');
legend('Location', 'eastoutside');
set(gca,'fontname','times')
ax = gca;
```

```
ax.XRuler.TickLabelInterpreter = 'tex';
ax.FontSize = 20;
ax.TickLength = [0.02,0] ;
ax.LineWidth = 2;
drawnow

end

%Function that forms the likelihood fucntion
function L_PH = like_PH(N,NH,PH)

L_PH = nchoosek(N,NH)*(PH)^NH*(1-PH)^(N-NH) ;

end


%Function that forms the triangular prior based on the bias

function p_PH = prior_P1_TRI_bia(PH,b)

if PH >= 0 && PH <= b

p_PH = 2/b*PH ;

elseif PH > b && PH <= 1

p_PH = 2/(1-b)*(1-PH) ;

end

end
```