Bayesian Inference vs Frequentist Inference: A Comparative Study

Introduction:

Statistical inference is an integral part of every researcher's toolkit and plays a crucial role in the development of scientific theories, Hypothesis, models, and their subsequent falsification, if necessary. Given the importance of statistical inference in scientific inquiry, it is paramount that we analyze the different approaches to statistical analysis of data, and their respective advantages & disadvantages. The dominant approach to statistical inference for the last 70 years has been the frequentist way of thinking, and the field of psychology and behavioural science is no different. Even though the frequentist approach of Hypothesis Testing is inundated with criticisms from several statisticians and experts in behavioural science (Kline. R, 2004), the use and misuse of p-values are prevalent. Kline has suggested several alternatives to NHST in his book, "Beyond Significance Testing" such as Robustness statistics using the bootstrap method, The "New" statistics of Confidence Intervals (CI), Effect sizes, meta-analysis, and the Bayesian approach.

In this article, we will mainly focus on the Bayesian approach, comparing and contrasting it with the classical frequentist inference. Throughout the essay, we will review and contrast the critical concepts of Frequentist inference and Bayesian inference. We will further identify contexts in which one deems better than the other by studying their respective advantages and disadvantages. Finally, we conclude with a hybrid solution from the bayesian and frequentist school of thought, which can act as a compromise between the two methods.

Foundation:

To explain the foundation of frequentist and Bayesian inference, we use the four-fold framework mentioned in Kruschke & Liddell (2017). This framework gives us a starting point to understand the different elements present in each approach thoroughly. To understand the framework, it is essential to grasp the idea behind what each cell of this 2×2 matrix refers to. Each cell reflects a different kind of analysis that a researcher wishes to do on a mathematical model of data. Let us briefly review what the previous sentence means.

Analysis of all kinds of data (observational, experimental, behavioural, neurological) begins with a formal description of the data using mathematical models. These mathematical models are made of meaningful parameters whose different values determine the data outputted by the model. Let us consider a problem statement to bolster our understanding. Consider a dataset of the height of 100 male students in ABC school. By plotting a histogram of this data, we may infer that a unimodal distribution like Normal distribution could be one plausible fit. The normal distribution has two parameters, namely mean (μ) and standard deviation (σ). By changing the values of these two parameters, we can produce data which resemble closely to the observed data. Once a model is chosen to describe the data, **what** we do with the model's parameters and **how** we do that constitute the four categories of the framework.

What/How	Frequentist	Bayesian
Hypothesis Test	p-value (Null Hypothesis Significance Testing)	Bayes Factor (Bayesian Hypothesis Testing)
Parameter Estimation	Maximum Likelihood Estimate with Confidence Interval	Posterior Distribution with Credible Interval

Table-1: Four-fold framework

If we intend to **examine** the parameter value of the model by a Hypothesis Test, we are concerned with the first row of the matrix whereas, if our goal is to **estimate** the parameter values, we move to the second row. "How" we do both these processes constitute the distinction in columns. Most of our essay focuses on the first row of the table, i.e. Hypothesis Testing and not parameter estimation. We will dissect the differences between Bayesian and Frequentist approaches by looking in detail about this 'How'.

Frequentist Approach:

Going back to our previous example of 100 male students' height, we may wish to examine the parameter of the Normal distribution by a Hypothesis test. In any Hypothesis test, we have a Null Hypothesis and an Alternate Hypothesis. The hypothesis framing depends on the research question, and in our case, we may assume that the research question is, "Is the male students' height of ABC school different from the average height of male students in the entire city?" We can define this formally as,

$$H_0: \mu = 175$$

$$H_1: \mu \neq 175$$

Once the research hypothesis is defined, we **reject** or **fail to reject** the Null Hypothesis based on the p-value. If the p-value is lesser than a certain threshold (usually 0.05), we reject the Null Hypothesis, else, we fail to reject it. So how is this p-value calculated? We calculate a summary statistic called t-statistic for the observed data. Then, we simulate many data sets similar to the observed data set (i.e. N = 100) under the assumption that H_0 is true (i.e. $\mu = 175$). Next, we calculate the t-statistic for each of those data sets and form a probability distribution of this t-statistic data (called sampling distribution). Under this distribution, the probability of finding a t-value as extreme as or more extreme than the observed t-value is defined as the p-value. If this probability is minimal, it means that the observed data is very unlikely under the Null Hypothesis assumption. Since we have observed the data empirically, our assumption of Null Hypothesis being true is wrong. Thus, p-value represents the **probability of getting a result as extreme or more extreme than the obtained result under the assumption that H0 is true**.

This definition of p-value holds for all distributions (t, F, Chi-squared). Still, it is essential to note that, for the same data set, different p-values can be obtained based on different sampling distributions. The sampling distribution is dependent on the stopping rule used by the researcher. Sanborn (2020) gives a perfect example to explain this phenomenon. Suppose that we have three heads in 12-coin tosses, but in one case, the researcher stopped after 12 tosses and, in another case, the researcher stopped after getting three heads. For the same data set, the first researcher will fail to reject the Null Hypothesis of the coin being fair, whereas the second researcher rejects the Null Hypothesis. The discrepancy arises due to different sampling distribution in each case (Normal distribution in the first case and Geometric distribution in the second) which leads to different p-values. This dependency on hypothetical sampling and the stopping intention of the researcher is considered as one of the drawbacks of p-value and NHST. We will discuss in detail about the other disadvantages of NHST in later sections.

Bayesian Approach:

Bayesian and Frequentist are two fundamentally different approaches in assigning a probability to an outcome. A frequentist assigns probability based on the ratio between the number of occurrences of favourable outcomes and the number of occurrences of all possible outcomes. In contrast, a Bayesian assigns probability based on his/her degree of belief on the favourable outcome's occurrence. Although this assignment of probability looks subjective, the core of Bayesian inference is the Bayes theorem which assigns a considerable role to the objective 'data' in the form of Likelihood. In a nutshell, Bayes theorem is all about having a prior belief about an event, observing the data, and updating this prior belief. Going back to our mathematical model of data, we might have a prior opinion about the parameter values of the model. The goal is to update this prior belief based on the observed data and form our posterior belief.

Formally, it can be described as:

$$P(\Theta|Data) = \frac{P(Data|\Theta) \times P(\Theta)}{P(Data)} \tag{1}$$

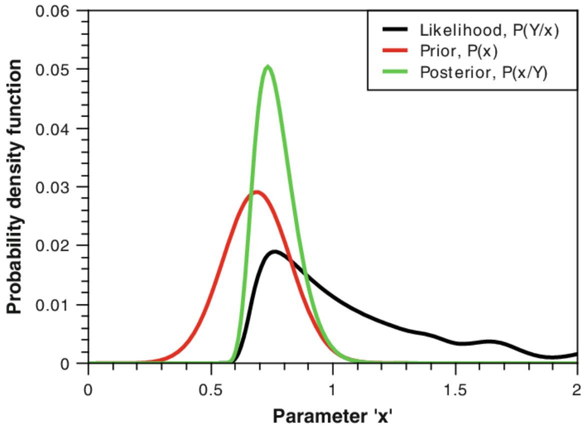

Figure-1: Prior distribution, Likelihood, and Posterior distribution.

The P(Data) in the denominator is a **normalizing factor** which shall be elaborated in later sections. In essence, Bayes theorem states that Posterior ∝ Prior × Likelihood. As we observe from the figure, if we are reasonably uncertain about the parameter values, we would have a flat prior distribution. Subsequently, the posterior will be determined by the Likelihood. The narrower the curve, the more sway it holds over the posterior. The critical difference between frequentist approach and the Bayesian approach is the additional factor of prior distribution which can moderate the effect of Likelihood (data). Thus, by multiplying the conditional probability of data given the parameter with the prior belief about the parameter, we can update our knowledge about the parameter value in light of new evidence. Now that we are done with the basics of bayesian inference let us see how this can help in Hypothesis testing.

Hypothesis testing is intuitive and relatively straightforward in the Bayesian setting. All we need to find is P(H_0|Data) which gives the probability that our Null Hypothesis is true, given the data. This value can be obtained through the Bayes formula by using the method mentioned below. We can use the same idea to find P(H_1|Data), the alternate hypothesis' probability.

$$P(H_0|Data) = \frac{P(Data|H_0) \times P(H_0)}{P(Data)} \tag{2}$$

However, the denominator in the above formula is relatively tedious to calculate, and this can be easily avoided by calculating the ratio between P(H_0|Data) and P(H_1|Data) making our inference comparative.

The final equation looks like this:

$$\frac{P(H_0|Data)}{P(H_1|Data)} = \frac{P(Data|H_0)}{P(Data|H_1)} \times \frac{P(H_0)}{P(H_1)} \tag{3}$$

The left-hand side of Equation-3 is called the posterior odds, while the second part in the right-hand side is called the prior odds. Prior odds represent the ratio between the probability of Null Hypothesis being true and the probability of Alternate Hypothesis being true before looking at the data and posterior odds represent the same thing after looking at the data. The middle part in the equation is known as the Bayes factor (BF_{01}), which represents the probability of Data under the Null Hypothesis relative to the Alternate. It is important to note that the Bayes factor makes no distinction between the Null and Alternate Hypothesis, i.e. it is just as easy to compute BF_{10} as BF_{01}. Thus, Bayes Factor is equal to Posterior Odds / Prior Odds, and if a researcher starts with a prior belief that H_0 is just as likely as H_1, then BF_{01} signifies how much more Null Hypothesis is plausible relative to alternate Hypothesis after observing the data. That is, $BF_{01} = 3$ for prior odds of 1:1 implies that the Null Hypothesis is three times more likely than the alternate Hypothesis after considering the data.

However, how do we calculate the Bayes Factor? The value P(Data|H_0) is called the Marginal Likelihood, and figuring it is non-trivial. In theory, we can calculate the marginal Likelihood of data given the Hypothesis by using Equation-1 in a slightly different context.

$$P(\Theta|Data, H_i) = \frac{P(Data|\Theta, H_i) \times P(\Theta|H_i)}{P(Data|H_i)} \tag{4}$$

The denominator in Equation-4 is what we require to calculate Bayes Factors, and this can be done by applying the general approach to find the denominator in Bayes rule. We mentioned earlier that the denominator represents a normalizing factor, and this factor can be calculated by summing the numerator obtained under different values of the conditioned variable. Assume that, in Equation-1, only three values for Θ are allowed. Then,

$$P(\text{Data}) = P(\Theta_1)\times P(\text{Data}|\Theta_1) + P(\Theta_2)\times P(\text{Data}|\Theta_2) + P(\Theta_3)\times P(\text{Data}|\Theta_3). \quad (5)$$

However, if the parameter space is continuous and not discrete, then we need to integrate over the entire parameter space. By extending Equation-5, we get that

$$P(Data|H_i) = \int_{\Theta} P(Data|\Theta, H_i) \times P(\Theta|H_i) d\Theta \quad (6)$$

It is crucial to note that if a model has more than one parameter, it becomes a higher dimensional integral, which is intricate to solve and finding the Bayes Factor becomes difficult. Undergraduates need not worry about this integration because it is calculable only for simple models or under special cases (like conjugate priors). Computational methods like Markov Chain Monte Carlo (MCMC) is the most preferred method than the analytical solution to calculate the posterior but explaining MCMC algorithms will take an article of its own and hence, we will not discuss it here. Still, MCMC become more challenging to implement as the complexity of the model increases (Lee and Wagenmakers, 2014). This challenge is considered as the computational problem in Bayesian Hypothesis testing by Lee and Wagenmakers (2014). The other problem is conceptual, and it is the fact that Marginal likelihoods are highly sensitive to the priors of the parameter distribution ($P(\Theta|H_i)$ inside the integral).

In Hypothesis testing, researchers generally have only a vague idea about the prior distribution of parameter values in H_1 (but the prior for H_0 is super specific) and using an uninformative prior lead to favouring the Null Hypothesis. More massive the prior width (i.e. more uninformative), more complex the model becomes. Increased complexity leads to the inclusion of unlikely values of a parameter, which decreases the overall Likelihood of the data under this model. Thus, using an informative and sensible prior is an integral part of the Bayes Factor calculation. It should also be checked if the Bayes factor varies considerably for a slight change in prior width through sensitivity analysis.

Thus, we looked at how Hypothesis testing can be carried out through the Bayes Factor, the process involved in its calculation, and some potential problems. In the next section, we will summarize the advantages and disadvantages of both the approaches from the understanding that we gained through the 'how' of bayesian and frequentist inference.

Advantages and Disadvantages:

To begin with, let us discuss the disadvantages of NHST in Hypothesis testing. We have already outlined how **stopping intentions** produce different p-values for the same data set due to the usage of 'imaginary' sampling distributions. This is not the case in Bayes factor (BF) because there is no sampling distribution involved in its calculation, and all the inference is derived solely on the observed data. Another drawback of the p-value is that, as the **sample size gets larger and larger,** H_0

is more likely to be rejected, and this incentivizes people to collect data until the p-value threshold is reached. This is referred to as 'p-hacking' in academic circles and to mitigate this, measures like pre-registration were introduced. This is not the case in Bayesian Hypothesis testing because BF results accurately reflect the data, and 'interpretation does not depend on the stopping rule' (Sanborn 2020).

If the p-value is significant, we reject H_0; else, no conclusion can be reached. The p-value does not say anything about **evidence in favour** of H_0. BF makes no such distinction between the Null and Alternate Hypothesis and as such, evidence in favour of H_0 can be obtained just as evidence opposing H_0. More importantly, the information provided by p-value **does not quantify** the evidence derived from the data. All we can infer is whether H_0 can be rejected or not. BF, on the other hand, quantifies the evidence and states precisely how likely H_0 is relative to H_1 in the light of observed data.

There is also the case of **'effect size fallacy'** where researchers misinterpret very low p-values (0.001) as strong evidence for H_1 compared to nominal p-values (0.04). The magnitude of the effect can be obtained only through effect size calculation and its corresponding confidence intervals in the frequentist setting. Kruschke and Liddell (2017) argue that this misinterpretation is common because what we get from NHST is not what a researcher wants. P-value does not give the **probability of the Null Hypothesis**, but often, this is what a researcher would like to know and hence interpret low p-values as strong evidence against H_0. This is in stark contrast to BF because it states explicitly what the probability of one Hypothesis is, relative to another. Thus, the Bayes Factor is intuitive and very easy to interpret.

Despite these advantages, the calculation of BF does suffer from problems of its own as discussed earlier. If a researcher has absolutely no idea about the prior of the parameter or if the complexity of the underlying model is high, then frequentist inference is a better alternative. This is because it does not require prior distribution to be stated and is computationally efficient for complex models. To circumvent these problems, Wagenmakers (2007) offers a hybrid solution which does Bayesian Hypothesis testing through BIC Approximation. This method is particularly useful when NHST yields an inconclusive result. In such a case, the publication should be based on BF value so that future meta-analysis does not get biased against non-significant results (Kruschke and Liddell, 2017).

BIC Approximation:

BIC has been used in model comparison of nested models for a long time, and it has been found that the difference in BIC value of two Hypothesis can be used to obtain Bayes Factor using a simple transformation.

$$BF_{01} = e^{\frac{(BIC_{H1} - BIC_{H0})}{2}} \qquad (7)$$

$$BIC(H_i) = -2 \times loglikelihood_i + k_i \times \log(n) \qquad (8)$$

Log-likelihood for Linear models like ANOVA can be calculated by dividing the Residual Sum of Squared (RSS) by the number of observations (n). k here refers to the number of free parameters in the model. Thus, BIC approximation takes the best out of both approaches and allows us to define BF using frequentist concepts like log-likelihood. The disadvantage in BIC approximation compared to the usual way of calculating BF is that this method uses ‘unit information prior’, which uses information only as much as the information in a single observation of data. This, in a sense, is less informative than usual and leads to bias against H_1. Thus, this method could act as a compromise - imperfect Bayes Factor, but can support NHST when inconclusive p-values are attained - resulting in less ‘publication bias’.

Conclusion:

To conclude, the Bayesian Hypothesis testing seems to provide answers to many of the criticisms of NHST and stands as a viable alternative to classical frequentist inference. The advantage of being unaffected by the stopping rule and the ease with which meta-analysis can be carried out in a bayesian setting promotes 'cumulative knowledge' (Ortega and Navarrete, 2016). In the present climate of replicability crisis, black and white thinking presented in the NHST framework will only limit the growth of the field and increase incoherence. The Bayesian inference could be the first step in moving away from this dichotomous thinking and in promoting the pursuit of cumulative knowledge in psychology.

With the advent of technology and increased computing power, software for bayesian inference are as widely available as standard packages used for frequentist analysis. However, care must be taken in Bayesian analysis for arbitrary priors only result in arbitrary BF values. Thus, if used in the right manner, BHT has more to offer than NHST and could also be used to bolster NHST when required.

References:

Kline B. Rex (2004), “Beyond Significance Testing: Reforming data analysis methods in Behavioral research”.

Krushke K. John, Torrin M. Liddell (2017), “The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective”

Lee MD, Wagenmakers E-J (2014), “Bayesian Cognitive Modeling: A Practical Course”

Sanborn Adam (2020) : https://my.wbs.ac.uk/$/$/$/event/cmsfile/t/item/i/1051202/v/1/f/0/n/week_6_bayesian_stats_v2.pdf

Wagenmakers E-J (2007), “A practical solution to the pervasive problems of p-values”

Ortega Alonso and Navarrete Gorka (2016), “Bayesian Hypothesis Testing: An alternative to Null Hypothesis Significance Testing (NHST) in Psychology and Social Science”