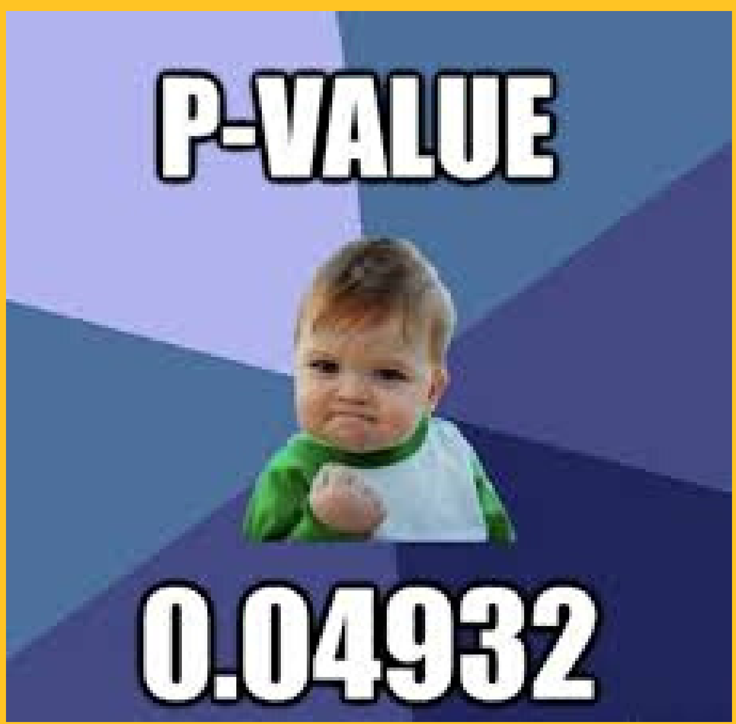
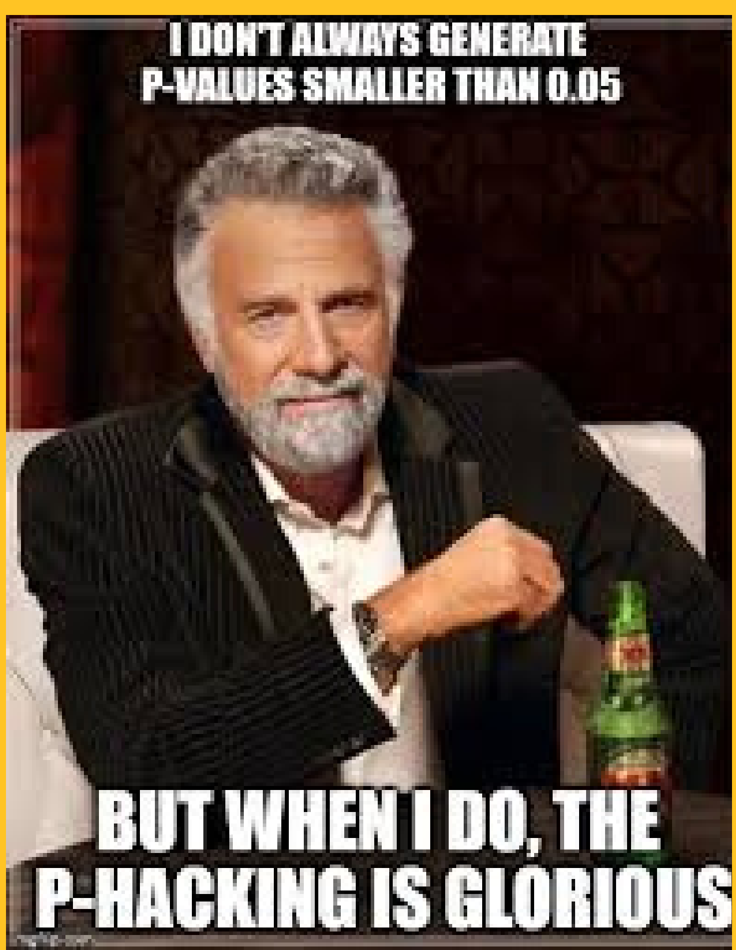
DID SOMEONE SAY BAYESIAN?

Undergraduate students have generally ignored Bayesian Statistics owing to its complex mathematics and the relative ease with which one can calculate 'p-values' from frequentist statistics. This short educational piece intends to shed light on the drawbacks of p-values, DOs and DONTs of its interpretation, and how Bayesian stats address these drawbacks. So, let us get cracking!!



- P-values **DO NOT** represent the probability of your Null Hypothesis (H0 from here on) but rather, the probability of your data. Common misinterpretation among students is that low p-value means the probability of my H0 being true is small. On the contrary, low p-value means, the chance of observing your data is small under the assumption that H0 is true. Complicated right? That's the p-value for you!! Unintuitive.
- **DIFFERENT** p-values can be obtained from the **SAME** data by various researchers. This phenomenon is called the 'stopping-intention' drawback of a p-value. Suppose that, I got three heads after 12 coin tosses and two researchers try to find the probability of the tossed coin being 'fair'. If one researcher assumes that I stopped the tossing experiment after 12 tosses, he will fail to reject H0. On the other hand, if the 2nd researcher assumes that I stopped the investigation after three heads, then he would reject H0. Thus, for the same data set, one can reach different conclusions about our poor coin in the frequentist setting.



- Also, p-values are highly sensitive to the sample size of underlying data with **bigger** sample sizes resulting in **smaller** p-values. Thus, a researcher can keep collecting data until he reaches the golden threshold of 0.05. This practice is called p-hacking in academia, and measures like preregistration of experiments have been suggested.
- P-value **DOES NOT** quantify the evidence obtained. What I mean by this statement is that a small p-value (0.0001) compared to a nominal p-value (0.03) does not imply that the former is strong evidence against H0 compared to the latter. All we can do with p-value is either reject or fail to reject H0 but cannot comment on the strength of obtained evidence against H0. This misinterpretation is so prevalent even among academics and is called 'effect-size fallacy'.

There are, of course, many criticisms of NHST cited by statisticians and experts. Still, the above four are the most important and relevant for Undergrads. So, how does Bayesian statistics overcome these drawbacks? In Bayesian Hypothesis Testing (BHT), we calculate something called **Bayes Factor** (BF). Unlike a p-value, BF directly gives us the relative probability between H0 and H1 given the observed data* and is thus intuitive and easy to interpret. Also, a BF of 3 means that given our observed data, H0 is three times more likely to be true than the alternate and thus, quantifies the evidence that we have. More importantly, BF does not depend on the stopping-intention of the researcher because unlike NHST, it does not depend on imaginary data** but only on the data, we observed. Thus, there is no room for discrepancies among researchers and results in uniform conclusions. Finally, BHT is not sensitive to the sample size of our study, and a researcher can collect data as long as one wants. This property makes them particularly useful in meta-analysis and in laying a foundation for cumulative knowledge***.

*: This is not entirely true, but will do for the time being. To make the article brief and non-technical, I am dropping the prior odds consideration.
**: By imaginary data, I mean the simulated data used to obtain the sampling distribution. For more detail, refer to the p-value comic.
***: By cumulative knowledge, I mean that evidence and data from previous studies can be incorporated in the present study to update our belief about the Null or Alternate being true.