Bayesian vs Frequentist: The Battle Over Statistics

By: Derek Albosta

1 Introduction

If you have ever taken a statistics in college, you probably remember fiddling with p-values, looking at confidence intervals, and building simple statistical models. What you have learned was only one way how to do statistical analysis, but there is another. Bayesian statistics has become a hot new field now due to the increased power of computing. It is a statistical method built on the basis of probability and is concerned with explaining relationships of data with probability distributions. This article will compare and contrast the two statistical methods. Using a dataset about exercise obtained on Kaggle, I will compare the results of t-test and linear regression using both of the methods.

2 Statistical Terms

Both statistical methods have a few terms and concepts that are critical in understanding how they work. I have defined them below:

2.1 Frequentist

significance testing and p-values

Significance testing is the core concept behind Frequentist statistics. It is the main way used to evaluate a model. Significance testing is concerned with rejecting the null hypothesis of a given type of test. The null hypothesis assumes that there is no significance with our given model. To do this, we calculate a statistic known as the p-value. This value represents the probability of obtaining a result that is more extreme than what is observed in the data. Getting a small p-value indicates that we have significant results and reject the "null hypothesis", claiming that our model has significant results.

However, there are many issues behind the way p-values work. To start, the definition of the p-value is difficult to conceptualize as to what it actually represents. Furthermore, p-values are liable to change based on the manor in which data is collected. For example, we can look to see whether or not a given coin is fair when we flip it. Our null hypothesis would be that the chance of flipping heads is 50%. When gathering data, the p-value of this significance test would change depending on if the experimental method was to flip a coin 10 times or to flip a coin until 10 heads results occurred. Based on data from these two separate experiments, we could potentially have conflicting results on whether or not to reject the null hypothesis with a significance threshold. Because of the p-value's dependence on testing methods, experimenters can manipulate their data collection methods in order to reach the common 0.05 significance threshold.

confidence intervals

Confidence intervals are a range of possible values for a predicted average that we cannot reject as being significant. This is generally defined as the 95% confidence interval. In theory, we can think of it as a range of values that if we were to repeat an experiment 100 times, 5 percent of the time our experiment will predict a mean outside of this range. These values are tied to the p-value. If the p-value is less than the 0.05 significance level, then the confidence interval should not contain 0. Because of its dependence on the p-value, confidence intervals have the same problems as the p-value.

2.2 Bayesian

intro to Bayesian logic

At a basic level, Bayesian statistical procedures form a mathematical tool set which allows for the application of probabilities to problems, and further allows for updates to those probabilities based on observation of new data. When we build a Bayesian model, we are concerned with a range of credible values that we want to predict rather than just a singular metric. We define these range of values under the 95% High Density Interval (HDI).

prior distributions

A Bayesian model requires some prior distribution to start from in order to begin updating. Generally, we want to start with a prior distribution that models the data as closely as possible. For instance, if we wanted to think about who would win our presidential election, we might have some idea in our head about the chances of each candidate has for winning. If we were then to look political polling sites, we can have a more informed guess as to who would win. The more informed our prior distribution is, the better it will fit to the actual model when we are constructing it.

posterior distributions

The posterior distribution is the result of updating the prior distribution with new observed data. This is done by means of shifting the probability distribution along the newly observed data. In a way, we can think of the posterior as a compromise between our posterior distribution we are updating on and a newly observed data. The posterior should lie somewhere in between those two probability distributions. After adding all data observations to a model, the resulting posterior distribution gives the credible range of values for the given parameter it is modeled on using the HDI.

3 Comparing the mean of two groups

3.1 Frequentist two group comparison

In statistics, it is common to try to gauge the effect that different groups have on a variable. This is commonly done in medical studies where researchers would compare a control group to a group under a medication under trial. In general, we are testing with the null hypothesis suggesting that there is no difference between the groups. Using our exercise dataset, we can see if gender affects the amount of calories burned. First, let us look at our distribution of the exercise variable by Gender:

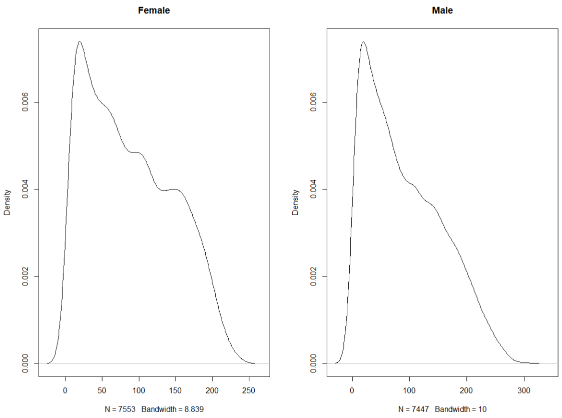
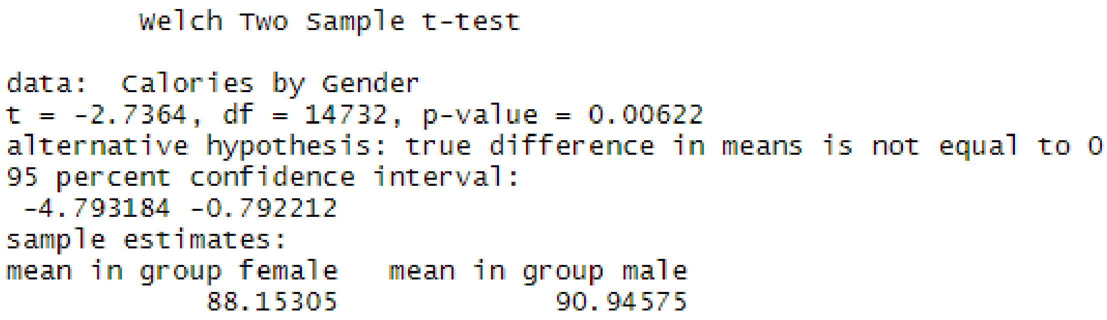


Figure 1. Density plot of calories burned on each Gender

Above, we can see that the exercise variable of both genders is not entirely normally distributed and has a slight skew. Let us conduct a Frequentist and Bayesian analysis on the data.

3.2 Frequentist two group comparison

We can perform a simple two sample t-test in R on gender and compare the means of calories burned for each group. Our null hypothesis states that the means of the two groups should be zero. Here is summary output from the Frequentist model:



```
         Welch Two Sample t-test

data:  Calories by Gender
t = -2.7364, df = 14732, p-value = 0.00622
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.793184 -0.792212
sample estimates:
mean in group female   mean in group male 
            88.15305             90.94575 
```

Figure 2. Results of conducting a Frequentist two sample t-test

In this t-test, the p-value represents the probability that the difference between the sample means is at least as large as what has been observed assuming that the population means are equal. A small p-value provides strong evidence that the two populations have different means. This t-test shows that we do indeed have a difference between the means since our p-value is below the 0.05 significance level. Our results indicate that there is around a 0.6% that the difference between the group means will be as large as what we have seen with our data. We would end up rejecting the null hypothesis and accept the alternative hypothesis.

The confidence interval of a t-test represents a set of limits in which you can expect the difference between the population means to lie. The p-value is less than the 0.05 significance level and we can see this is true with our results: our reported 95% confidence interval suggests that the difference of the means of calories burned when comparing female to males are around -4.79 to -0.792.

3.3 Bayesian two group comparison

We can create a simple Bayesian model over our two gender groups and compare the results of our posterior distributions. When we explored the data, we noticed that the data was not normally distributed. In a Bayesian model, we can account for that by choosing any type of distribution to represent the data. To give our model robustness, we will use a t-distribution. A t-distribution is just like a normal distribution, but with an additional parameter ν. This parameter add weights to the tail end of the distribution. The larger the ν is, the t-distribution will become more like a normal distribution. Using a t-distribution makes our model robust since it can help account for outliers in our data.

When choosing a prior distribution for this type of data, we can use the mean and standard deviation of our data. By setting our μ equal to the mean of each group and our σ to the standard deviation, we can get extremely close to the t-distributions actual parameter values. This would be considered an informed prior since we are using the data to help us converge to the true parameter values. The only variables we have not set is our ν. Since we have no way of interpreting what this value will look like based on previous knowledge, we can set it to some vague value. A ν parameter is generally sampled from an exponential distribution with an additional 1 added to it. In this model, an arbitrarily small value since the data was not normally distributed. Below are the posterior distributions of both gender' burned calories, the posterior in our nu parameter (log transformed), and their comparisons.

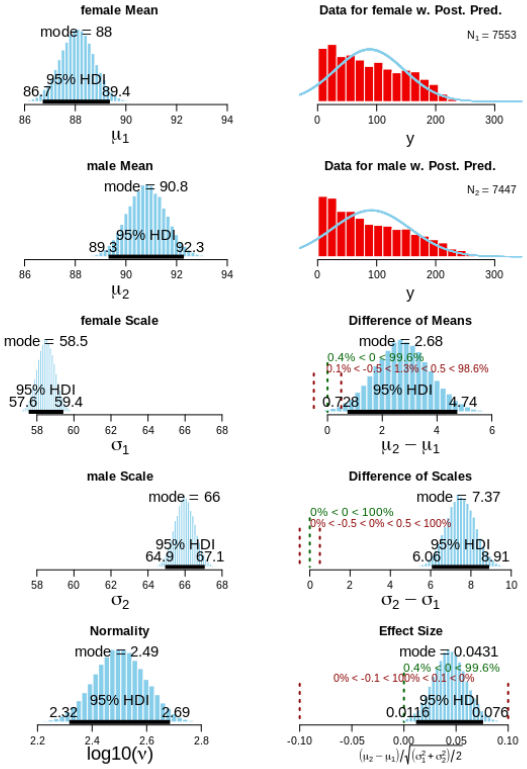


Figure 3. Results of conducting a Bayesian two sample t-test

The graphs on the left represent the posterior distributions of our parameters. We can see that the mean for females is most credible at 88 while for males it is 90.8 calories burned. Notice that the graphs display a 95% HDI range. This range represents where 95% of the data lies and can be considered the range of credible values that the actual mean of these groups can be. For females, this ranges from 86.7 to 89.4 and males its is 89.3 top 92.3 calories burned. Since the HDI range for males is larger than females, we can be a little more confident about our calculated average for females. This can be seen within our σ parameter where our predicted parameter for females are indeed lower than male. The ν parameter is shared by both distributions. Given that we have a small ν, this model ended up accounting for outliers. The top two graphs in the left corner represents the proposed distribution over our data. Notice how the tails are quite wide as a consequence of our small ν value. When comparing the groups, we are using ROPE to signify a range of where we can reject the null hypothesis. While looking at the difference of means plot, we notice that the HDI range does not fall within our ROPE bounds, thus we would reject the null hypothesis and state that the means are different with great certainty.

3.4 Comparison

Comparing the two models, our Bayesian model gives us a lot more useful information while also being adaptable to the shape of the data. The Frequentist t-test only gives us the ability to claim that there is a difference in the means. Our Bayesian model shows us this, but also shows other probable values that our groups' means can be. The confidence interval can only tell us that 95% of the time this experiment is performed, the calculated difference of the means will fall between -4.79 and -0.79. This is in contrast to our Bayesian calculated difference of the means, which gave us a range of credible values that the difference can be. One last difference, is that the Frequentist t-test always assumes data normality while the Bayesian model used a robust t-distribution to account for outliers. Either way, it is easier to interrupt and use the results of the Frequentest t-test since its hard to use a range of parameter values to predict anything meaningful.

4 Fitting a linear regression

Linear regressions are a common to model the relationship between two variables of data. This involves creating a line of best fit as a way to predict one variable with another, minimizing the total distance of the line and all of the data points. A linear regression model is defined as $\hat{y} = \beta_0 + \beta_1 x$ where $\hat{y}$ is the variable we want to predict, β_0 is the y-intercept, and β_1 is the slope multiplied by a given x predictor variable. Using the exercise dataset, we can attempt to create a linear model describing the relationship that exercise time has on burning calories. Let us first explore our data by plotting the two variables against each other:

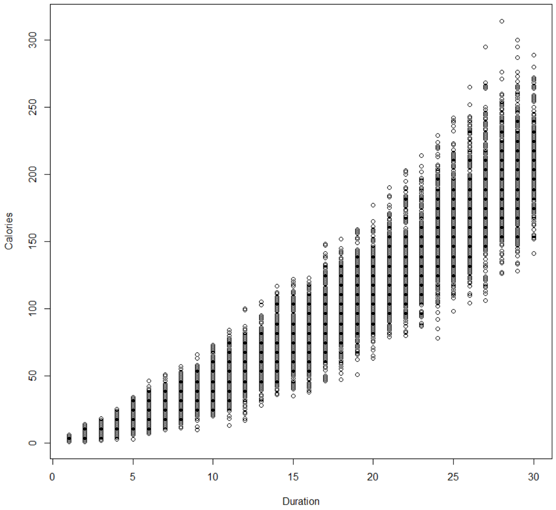


Figure 4. Plot of calories burned by exercise duration(min)

By looking at the data, it seems like there is a linear relationship between our variables. However, it is important to note that there seems to be increasing variance in Calories as Duration time increased.

4.1 Frequentist linear regression

We can first create a Frequentist linear model over the data to find a proposed y-intercept and slope. This method uses the mean squared error equation in order to reduce the amount of variance that our model generates. Least squares penalizes outliers heavily by squaring its distance from the proposed model line. Our null hypothesis states that there should be no relationship between the data (the slope = 0). Below is the proposed line of best fit shown over the data with its corresponding summary statistics.

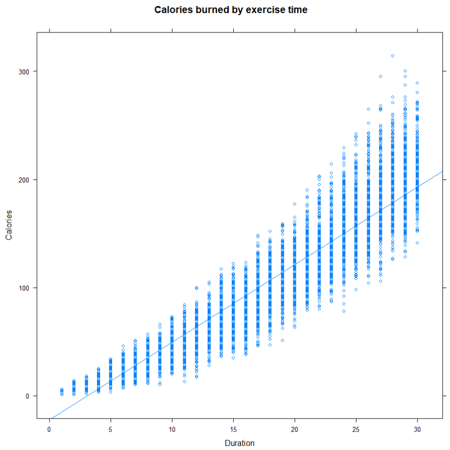
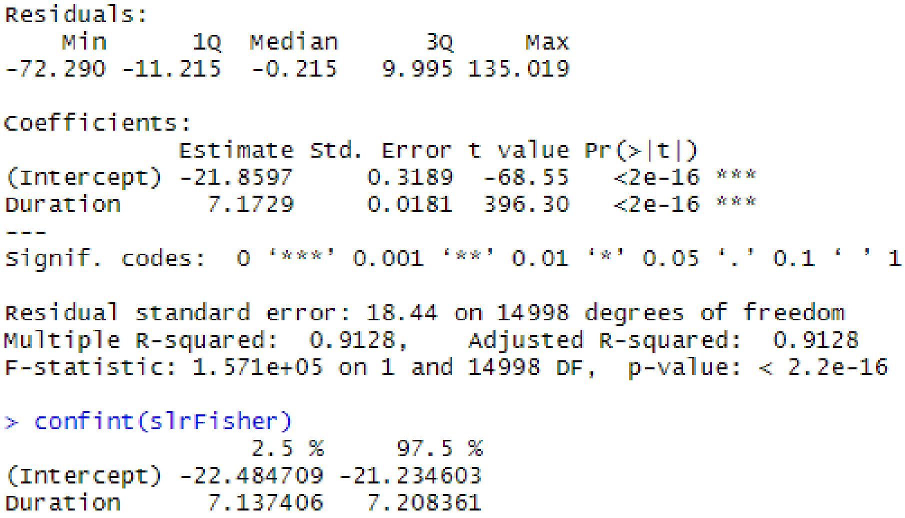


Figure 5. Mean squared error best fit line over data



```
Residuals:
    Min      1Q  Median      3Q     Max 
-72.290 -11.215  -0.215   9.995 135.019 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -21.8597     0.3189  -68.55   <2e-16 ***
Duration      7.1729     0.0181  396.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.44 on 14998 degrees of freedom
Multiple R-squared:  0.9128,	Adjusted R-squared:  0.9128 
F-statistic: 1.571e+05 on 1 and 14998 DF,  p-value: < 2.2e-16

> confint(slrFisher)
                 2.5 %     97.5 %
(Intercept) -22.484709 -21.234603
Duration      7.137406   7.208361
```

Figure 6. Diagnostic statistics for Frequentist model

This type of model generates two types of p-values: p-values for the parameters, and p-values for the model. If our p-value is significant at the 0.05 level, we will reject the null hypothesis (our variables are not significant). Our (Intercept) coefficient represents the y-intercept and the Duration represents the slope. Since our p-values are below the significance level, we can reject the null hypothesis and state that these variables are significant. Our p-value for the overall model is the same as our p-value for our intercept.

The confidence intervals for our coefficient variables represent the range of values in which there is a 95% chance that if the experiment was repeated, the true value would be found. We can see this range for our intercept runs from around -22.48 to -21.23 and our slope's range is from around 7.14 to 7.21.

The multiple R-squared value represents how well the model explains the variability of the data. This uses the average distance between the observed data and the linear model using the mean squared error to calculate the distance. Using this score, our model is said to explain about 91% of the variance.

4.2 Bayesian linear regression

We can create a Bayesian linear regression model to predict Calories burned with exercise duration. In a Bayesian linear regression, we can explain the variance of our data along our proposed linear model by using any type of distribution. We can again use a t-distribution to represent our parameters and give our model more resistance to outliers.

When choosing a prior distribution for this type of data, we do not have any good prior knowledge on how the our slope and intercept parameters can be represented. For this model, the prior for our slope and intercept is sampled from a normal distribution with vague priors. It is important to note though that for a Bayesian linear regression, we standardize the scale of our data to help the computer calculate the parameters easier. So, our prior distributions are set to a standardized scale and then converted back once the computer is done calculating. Our ν and σ parameters are also set with using vague priors. Below is our proposed Bayesian linear model with t-distributions fitted to represent the variance as well as the posterior distributions of our slope, intercept, σ and ν parameters. It should be noted that the plot of the linear regression actually has a multitude of regression lines plotted over the data with parameter values sampled from the generated posterior distribution:

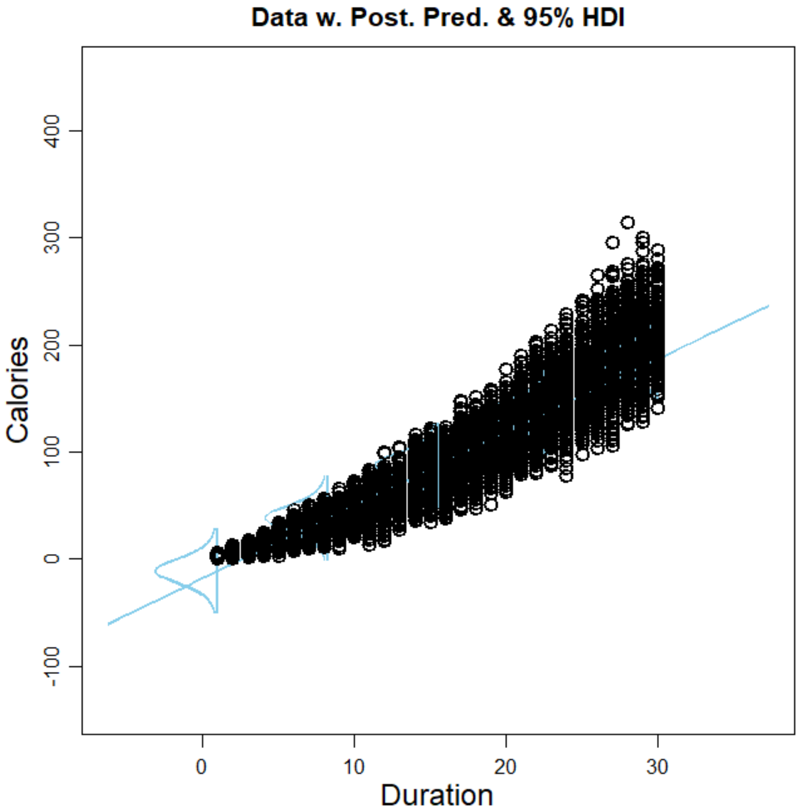


Figure 7. Prospective linear model with t-distributions for variance

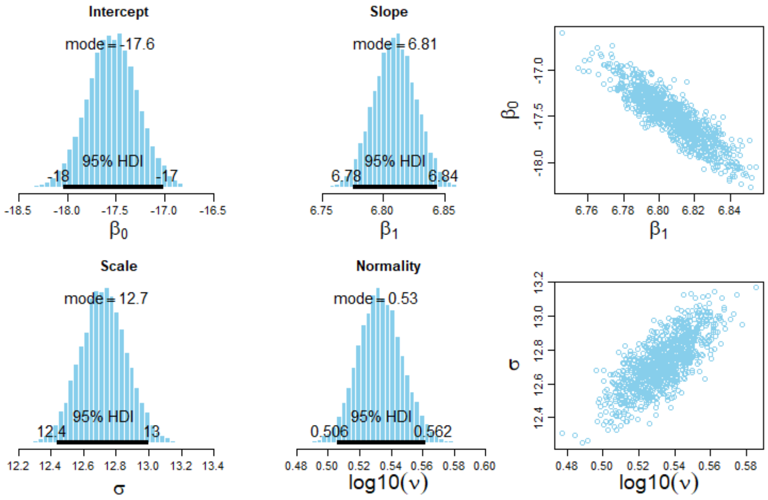


Figure 8. Bayesian posterior of parameters of linear regression

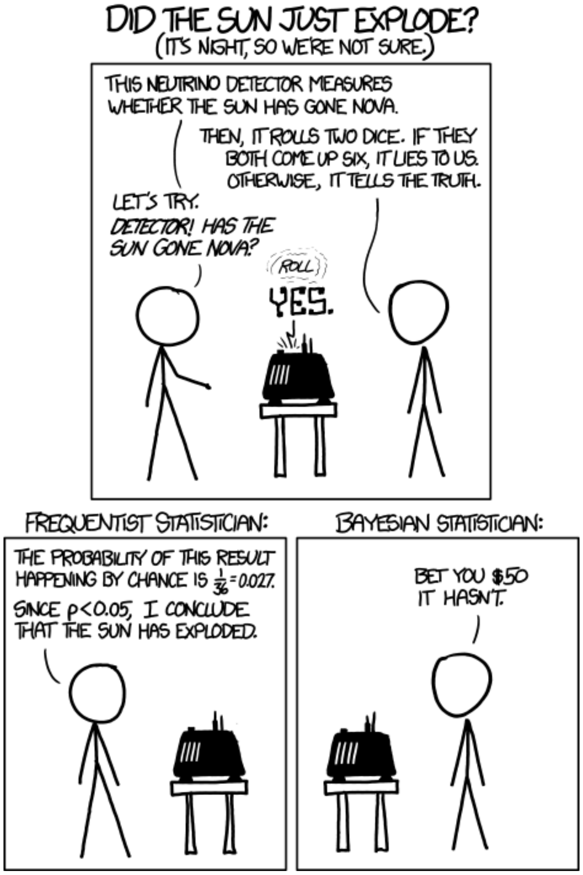
From our simulation, we are given a set of credible values for each of our models parameters within the 95% HDI. Our proposed y-intercept value is credible between -18 and 17 with the most credible value being -17.6. Our slope is credible in the ranges of 6.78 to 6.84, being most credible at the value 6.81. Our σ (scale) parameter is large and suggests that there is a lot of variance to our data. Our small ν parameter suggests that we are heavily accounting for outliers in our data.

4.3 Comparison

Between the two models, the results of the Bayesian model tell us more about the relationship between our two variables. It tells us more information about the variance of the model as well as how the outliers affect it. It is interesting to notice how the intercept we predicted for our Frequentist model much lower than that of our Bayesian. The slope for our Frequentist model is also a larger value. This is due to the assumed normality of the data in the Frequentist model and it is being affected by the outliers more. Again, if we are just trying to predict the amount of calories burned given how long someone exercised, it is harder to interpret the results for the Bayesian model since we are given a range of credible values.

5 Conclusion

To summarize, Frequentist statistics is concerned in absolutes while Bayesian statistics is concerned with credible probability. This is why we can create robust Bayesian models that fit well to many types of data while Frequentist models are reliant on satisfying assumptions about the data and creating significant summary statistic values. The burden of Frequentist statistics is how p-values and confidence intervals can change based on testing practices. Bayesian models give us more information about the relationships between our data with the use of parameter estimation and HDI ranges. In terms of practicality, Bayesian models are hard to use in the real world because we cannot create predictive models based on a range of parameter values. This is why applied statistics mainly used Frequentist methods in areas of machine learning because such algorithms are concerned with calculating exact values. To end, below is a funny comic that summarizes the difference of Frequentist and Bayesian interpretations of events.



DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.