Statistical Inference

Capstone Project # 1 - Predicting Popularity of YouTube Videos

Statistical Inferences Code

https://github.com/fariha23/YouTube_Data_Analysis_Video_Categories

Statistical Inferences Scenarios

Scenario 1:

For YouTube (Client) to consider investing properly in people, infrastructure etc, they need to know audience engagement with the videos in both countries. For this they need to know the views distribution and a 95% confidence interval around the population mean.

1. Considering data frame 'df_usa_views' to be a sample of all the video views in the USA, the goal is to infer population statistics for all the video views across all categories in the USA.
2. Considering data frame 'df_mx_views' to be a sample of all the video views in Mexico, the goal is to infer population statistics for all the video views across all categories in Mexico.

Approach:

1. For the USA and Mexico sample: Find sample statistics (mean, median and 95 percentile range of the sample views). Plot a histogram to visualize the data
2. For the USA and Mexico sample: Apply bootstrap approach to infer population parameters. Create 10,000 bootstrap samples from the given sample data and calculate the mean of each bootstrap sample.

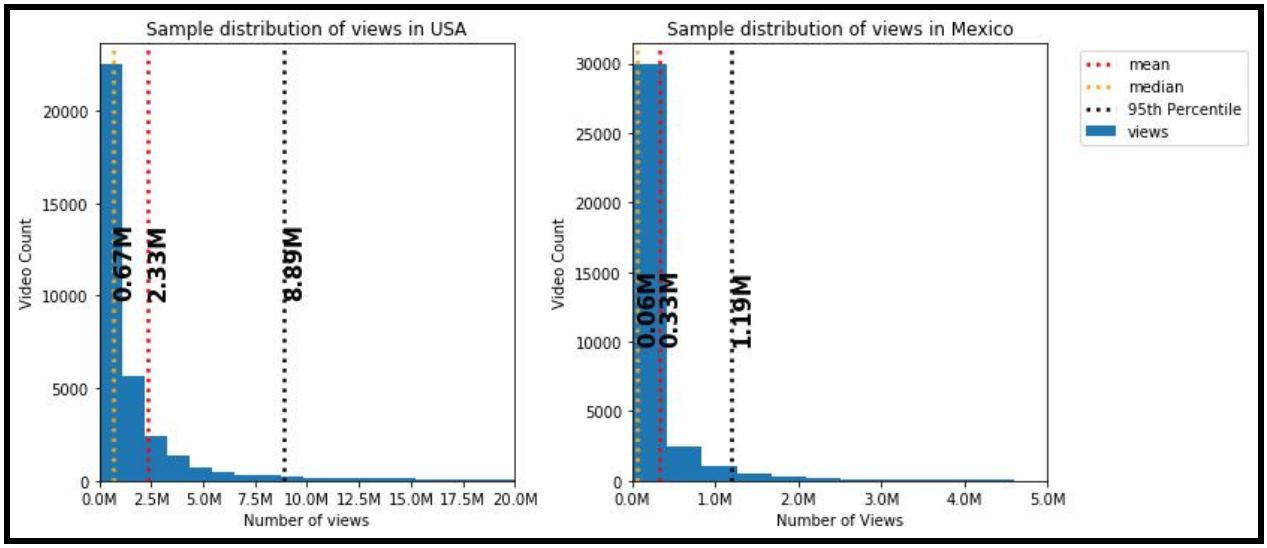

Figure 1: Sample Statistics - USA/Mexico

SUMMARY OF STATISTICS:

USA SAMPLE STATISTICS - VIEWS

.95% of the views are under: 8.89M
.Other statistics of the views in the USA:
.Sample Mean= 2.33M
.Sample Standard Error = 7.26M
.Sample Median= 0.67M
.The PEAK of the distribution is closer to it's median 0.67M, rather than mean 2.33M
...This is expected behavior from a left skewed distribution

Mexico SAMPLE STATISTICS - VIEWS

.95% of the views are under: 1.19M
.Other statistics of the views in the Mexico:
.Sample Mean= 0.33M
.Sample Standard Error = 1.45M
.Sample Median= 0.06M

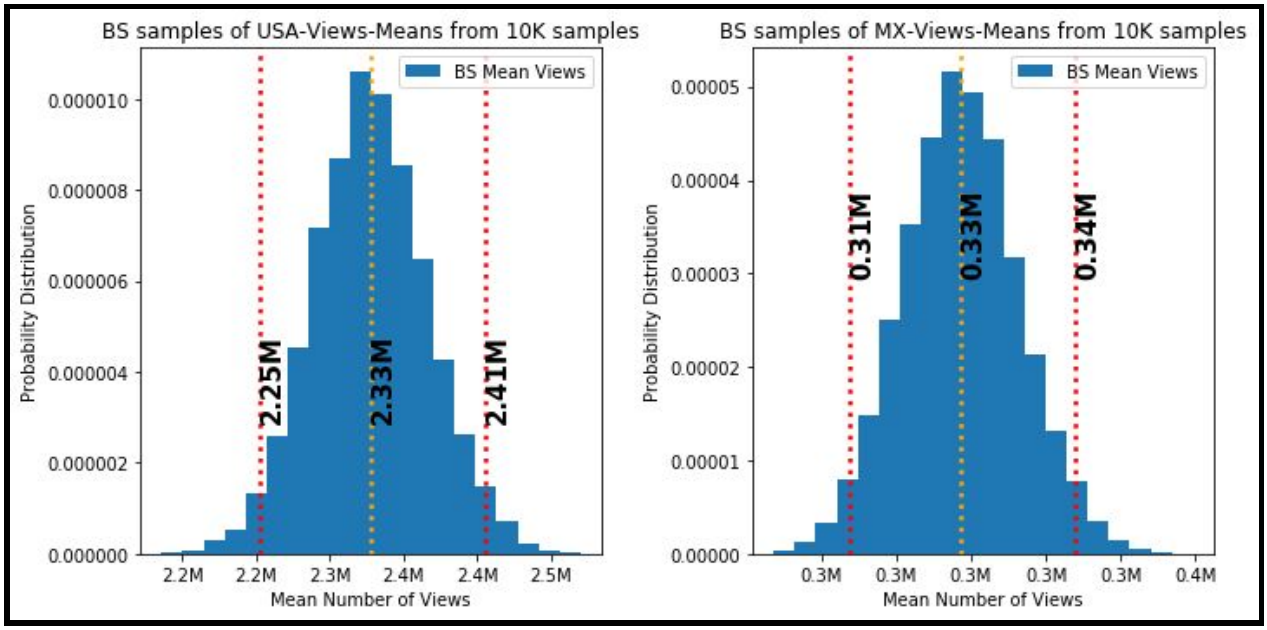

Figure 2: Bootstrap approach - Inference of Population mean with 95% CI

USA:Confidence interval (95%) for population mean is [2.25M ,2.41M]
Mexico:Confidence interval (95%) for population mean is [0.31M ,0.34M]

Analysis of Scenario 1

For USA-Views: With the bootstrap inference, it can be concluded that there is a 95% chance that the population mean (all videos views) is between 2.25M and 2.41M.

For Mx-Views: With the bootstrap inference, it can be concluded that there is 95% chance that the population mean (all videos views) is between 0.31M and 0.34M

From Figure 1, the histogram of views in the USA, it can be seen that the mean views of the sample is not close to the datapoint where the majority of the data lay. This is because the outlier (max views of one of the music category's videos 'Childish Gambino-This is America" has 217M) skewed the data. This anomaly is not really an anomaly since every year there will be video/videos more popular than others. Therefore by using bootstrap method for inference an average number of views in the USA and MEXICO was determined without having the outlier's influence on the statistics (as shown in Figure 2)

Scenario 2:

Does the number of videos in each category depend upon the country it is viewed in? or country has no effect on the video categories and the number of videos in each category?

Approach:

Using the Pearson's chi-squared statistical hypothesis test to determine if there is dependence between the variables. For analysis imported chi2_contigency and chi2 methods from scipy.stats library. Using chi2_contigency method, first stat, p-value, degree of freedom(df) and expected value was calculated for the contingency table. The contingency table (required for chi squared test) was formed as below:

```
chi_sq_table_views=pd.pivot_table(dw.combined_usa_mx_df,index=["country"],
columns=['category_name'], values=['views'], aggfunc='count', fill_value=0)

#H0: We will assume that country and the category variables are independent (TRUE)

#H1: The country and the category variables are not independent
```

Note: Chi2 Test and it's results' interpretation and how is it used with Python was learned from https://machinelearningmastery.com/chi-squared-test-for-machine-learning/

Results:

Probability=0.95, Critical=25.00, Stat=7532.27
Stat is greater than critical value therefore variables are Dependent (reject H0)
Significance=0.05, p=0.00
P-value is less than alpha (significance level) therefore variables are Dependent (reject H0)

Analysis of Scenario2

Since results show that stat returned from the chi2-test is greater than the critical value, the two variables are dependent

Also the p-value returned from the chi2- test is less than alpha showing that the two variables are dependent

Scenario 3:

How does the expected value of the video views change as time progresses?YouTube could be interested in this data to see how and when to place advertisements and attract traffic to click on a new ad.

Let's say a new video which has the same genre and predictors as one of the videos in the sample, could using the inference data of the views of the sample video, be beneficial to the producers in determining when to expect an increase in the views after launching it

Part 1: The most popular video in the USA on YouTube is in the Music Category titled "Childish Gambino - This is America". It was introduced on May 8th 2018 and collected views in abundance till it reached 225M on June 2nd 2018. When did the video actually take off? What is **expected value of the views** from the day it started to the last day the data was collected for this YouTube data set

Part 2: What is **expected value of the difference in the views** (How many more or less views did the video receive everyday compared to previous day)

Approach Part 1: Expected value of the views

Using Bayesian Inference methods, an analysis is presented on the expected number of views collected between 2 periods (there could be more but for this analysis only two are presented). First the progress of views per day is determined and the data (bar chart) is plotted with the mean value of the views.

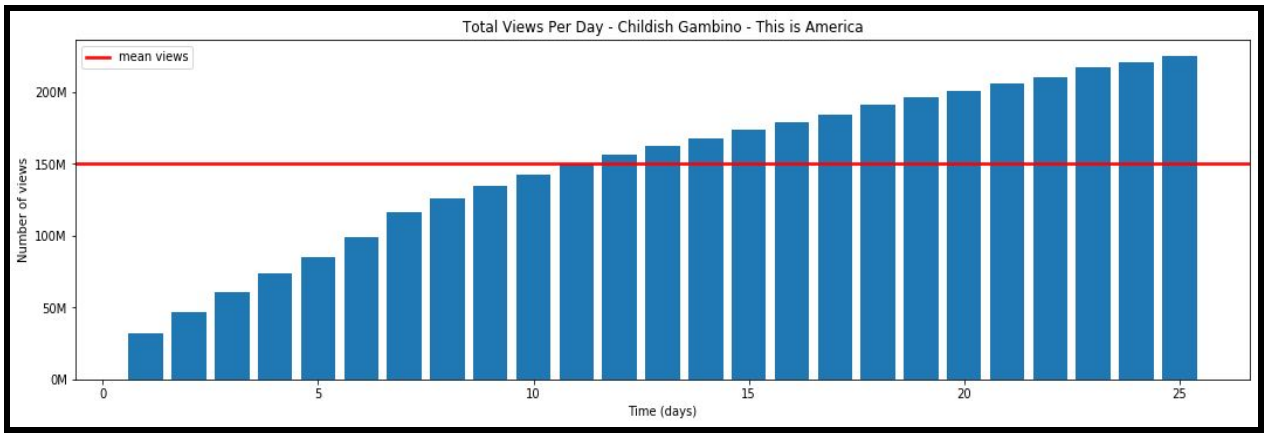

Figure 3: Views per Day - Childish Gambino - This is America

Observations from the views bar plot of Childish Gambino's - This is America Video Views

1. It can be observed from Figure 3, that the mean of several periods is not the same number of 150M.
2. From Baysian methods for the hackers book, we can start to model this by defining a Poisson Random Variable as it can model a count data. Therefore views count on day i ~ Poi(lambda)
3. Lambda is unknown
4. It can be observed from the bar chart in Figure 3 that the rate increases as the days go up.
5. However the rate is different for *at least* 2 different periods of times (days)
6. lamda_1 and lambda_2 are defined as those two periods.

7. And tau_1 is defined as a day when the change of rate occurs. There are several possibilities but keeping it as Uniform Distribution between 0 and 25 [tau_1 ~ DiscreteUniform(1,25)]
8. All the a prior distributions are defined and all the posterior distributions of lambda_1, lambda_2 and tau_1 are calculated using pymc3 module of python

Observations from Posterior Distribution (Figure 4)

1. Lambda_1 and Lambda_2 distributions are not exponential as we started with. The MCMC "engine" plugs in those two variables along with tau_1 and produces a posterior distributions that are more appropriate to the given views data of the video
2. Tau_1 distribution started with Uniform distribution and the MCMC determined that the change likely happened between 5th and 6th day of the release of the video

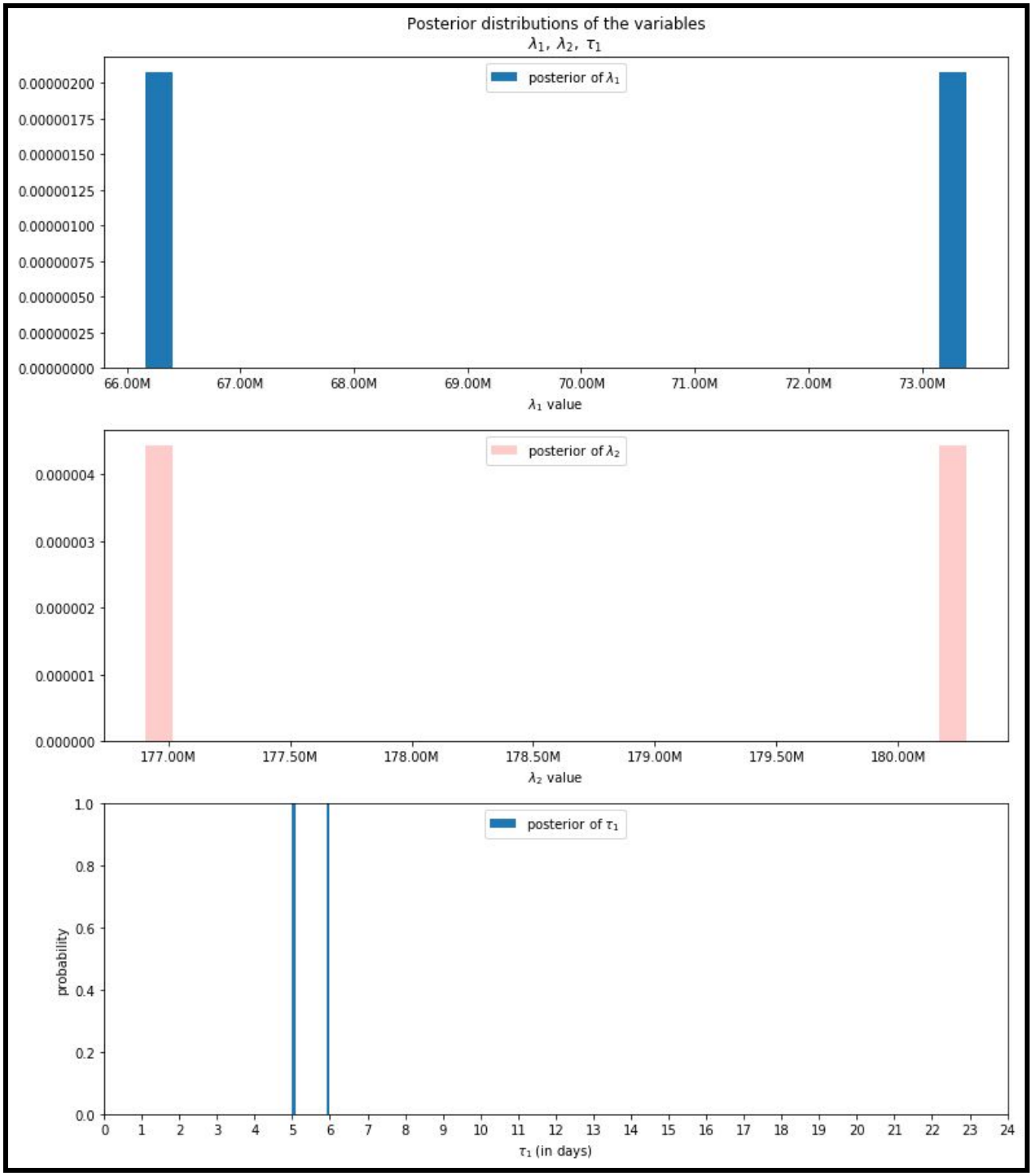

Figure 4: Posterior distributions of lambda_1, lamdba_2 and tau_1 - Views

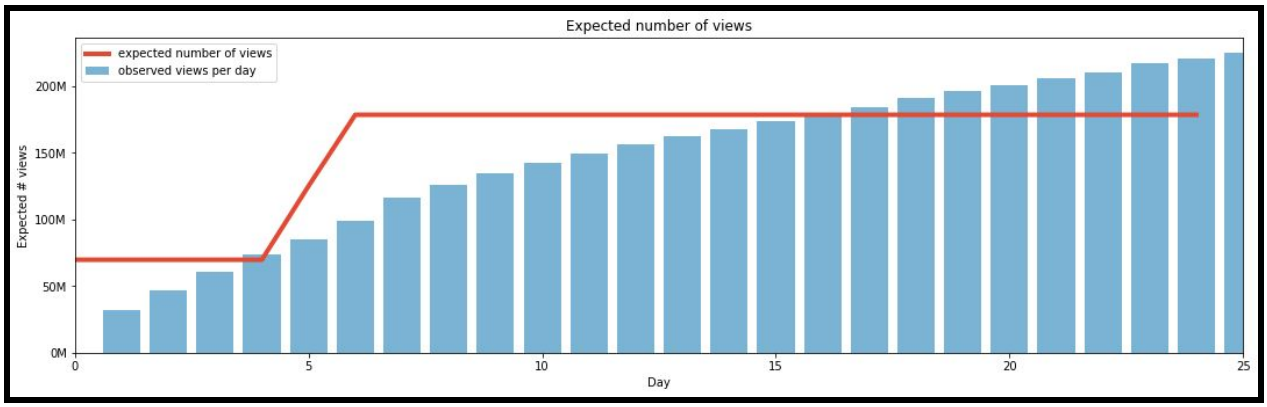

Figure 5: Expected number of views per day using Bayesian Inference - Childish Gambino

Analysis of the expected views graph on the switch points:

Clearly the change in views happened between day 5 and 6 (switchpoint). Expected value of views went from 75M to 180M views as shown by the posterior distributions at that switch point

This view is important as it gives more insight into how the mean of the views is changing from day one. The raw mean of the views from day 1 to 25 as shown in Figure 3 is 150M. This is not insignificant statistic however by showing means in the periods where video was just introduced vs when it actually took off gives more insight into the data

Approach Part 2: Expected value of the difference in the views

Using Bayesian Inference methods, an analysis is presented on the expected number of the difference in the views collected between 3 periods (there could be more but for this analysis only three are presented). First the progress of difference in the views per day is determined and the data (bar chart) is plotted with the mean value of the difference.

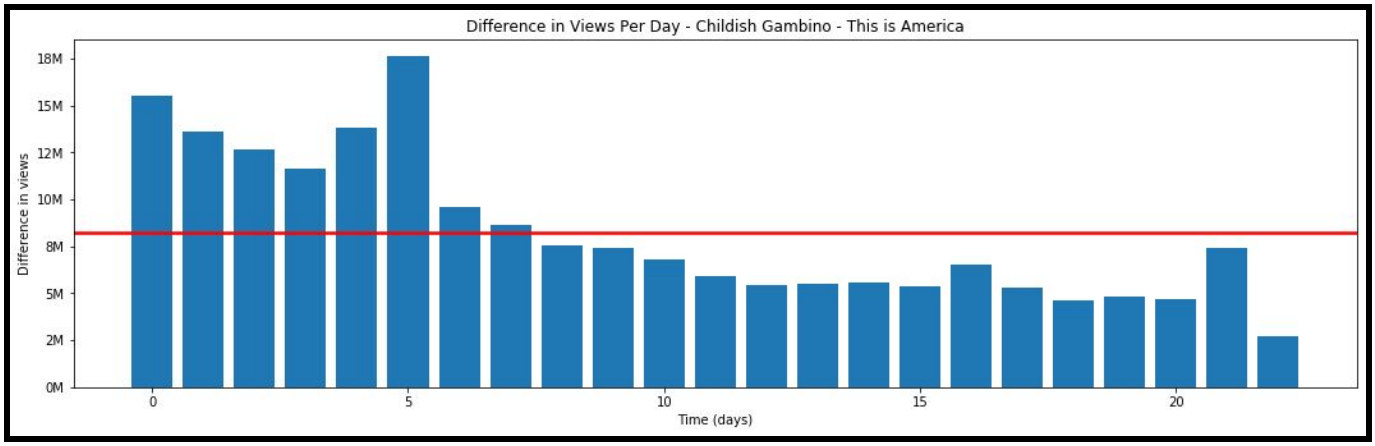

Figure 6: Difference in Views per Day - Childish Gambino - This is America

Observations from the difference in views bar plot of Childish Gambino's - This is America Video Views

1. The most difference in views occurred on day 5 which was predicted by our model previously i.e it chose day 5/6 accurately to predict the one switch point. It can be seen that there are more than 1 switch points from Figure 6 but since the previous model (Expected views per day model) only had one (tau_1) the MCMC engine chose day 5/6 as the most "viable" candidate
2. The mean of the difference in views is 8.2M
3. A similar model for difference in views is produced using pymc3.
4. Two switch points are chosen for this analysis however there clearly are more than two here. To keep complexity in control, two switch points are OK! Two switch points mean three different rates for three periods (lambda_1, lambda_2 and lambda_3).
5. Random variables lambda_1, lambda_2 and lambda_3 are defined as having exponential distribution
6. Random variables for switch points are tau_1 and tau_2 and are defined as having uniform distribution

Observations from Posterior Distribution (Figure 7)

1. The posterior distributions for random variables are not same as their a priori distributions
2. Note: Everytime the pymc3 model is run it tries to best fit the data to determine the posterior distributions. For this report Figure 7 posterior distributions are chosen however after running the model many times, the posterior distributions come up different every time.
3. For these posterior distributions, there are two switch points (Day 6 and Day 21). Other posterior distributions that were seen produced three possible switch point (Day 5/6 , 10 and 21)

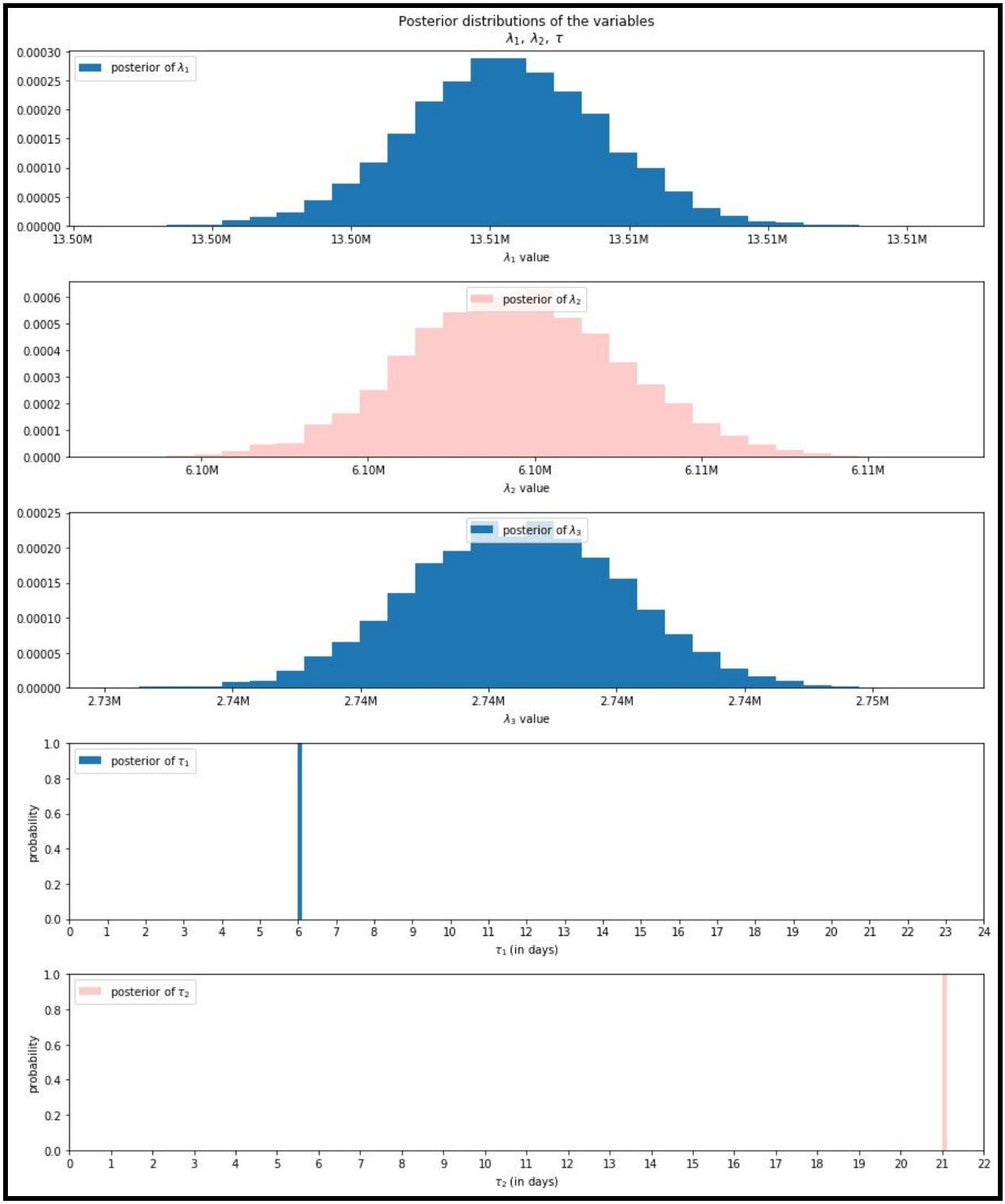

Figure 7: Posterior distributions of lambda_1, lamdba_2 and tau_1 - Difference in Views

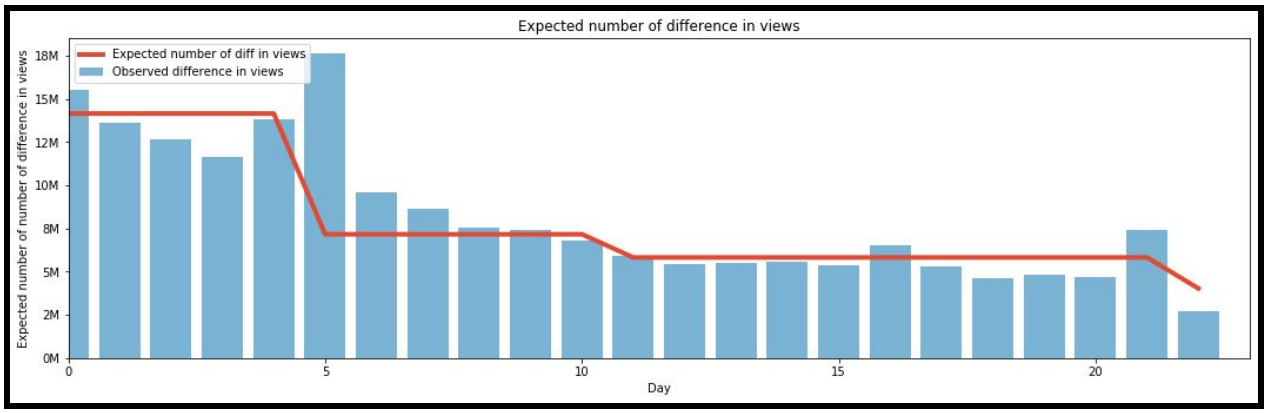

Figure 8: Expected number of difference in views per day using Bayesian Inference - Childish Gambino

Analysis of the expected number of difference in views graph on the switch points:

The model predicted that there are two dips in the expected values. Around day5 and day 21. As any popular video eventually starts declining in viewership as time passes, this model gives a glimpse of when the most views can be achieved and when the decline starts.