(), , 1–14

### ARTICLE

MEANINGFUL DECISION MAKING WITH BAYESIAN ANALYSIS

Halilou Junior Tmite, Tangbufan Wei, Gaurav Daven

**Abstract**

The two main approaches to statistical inference: Bayesian vs frequentist analysis have long been a topic of debate for statisticians. In essence, the frequentist approach deals with long-run probabilities whereas the Bayesian approach is concerned with the probability of a hypothesis given a particular dataset. One of the main strengths and possibly topic of concern for Bayesian analysis is that it incorporates prior information (based on some belief) into the analysis. On the other hand, the frequentist approach is only driven by the data at hand.

In this report, we aim to show two applications of Bayesian analysis: AB testing and Bayesian linear regression. We show through results obtained that Bayesian analysis is straightforward to interpret and can help with decision making.

**DATA SET**

ab˙data

**DESCRIPTION**

A/B testing experiment concerning two different landing page layouts in January 2017

mimic3d

Dataset comprising of deidentified

health-related data (demographics, lab results, procedures, etc.) associated with patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012

patients

Randomly generated dataset to predict the

length of stay (days) in hospital using the regression model built from the mimic3d

dataset

Keywords: Prior, Posterior, Likelihood

# INTRODUCTION

The usefulness of Bayesian analysis can perhaps be understood by the fact that it allows us to define a ”probability” in the same way that most normal people do, by indicating the plausibility of a proposition or situation. It allows us directly assign probabilities which then describe the

plausibilities of the possible outcomes for that particular situation based on some prior belief. Bayesian analysis allows us to calculate the probability that a certain hypothesis is true whereas frequentist analysis only allows us to calculate the probability of obtaining another dataset atleast as extreme as the one we have (through the p-value). Recently, Bayesian analysis is gaining popularity in the fields of medicine, engineering and law since it allows for a more intuitive interpretation of results. For the frequentist approach, the p-value obtained is often misinterpreted and hard to make decisions with in the real world.

In this report, **Section 1: Website Landing Page Analysis** is about the Bayesian AB testing for a website landing page. First, exploratory data analysis is conducted to understand the data visually and manipulate it for analysis. Data is divided into two groups: **Control (Old Page)** and **Treatment (New Page)**. Based on a set prior, we calculate the posterior probability concerning which page has a better conversion rate. **Section 2: Predicting Length of Stay in Hospital** concerns the dataset ”mimic3d” which is used to predict length of stay in a hospital given a wide variety of predictor variables such as age, martial status, insurance type, ethnicity, etc. Exploratory data analysis is used to determine which predictor variables might have a strong link to the response. Bayesian linear regression is then used to build a model using the ”mimic3d” dataset. Prediction of length of stay in hospital is conducted using the ”patients” dataset.

This report is also accompanied by data dictionaries (before the Exploratory Analysis section) to aid readers’ understanding of important variables in both datasets.

# WEBSITE LANDING PAGE ANALYSIS

**1**

This dataset contains the results from AB testing experiment for different versions of a website landing page. The data dictionary below (Table 1) is intended to help the reader understand the different variables.

Table 1. Data Dictionary for the AB Testing Dataset

**Variable NAME DEFINITION CODES DATA TYPE**

**user˙id** Unique ID of the user 290584 Distinct User ID. Numeric

**timestamp** Recorded time when the user visit the landing page.

January 2nd, 2017 to January 24th, 2016

Date

**group** User belonging Group. control / treatment Factor

**landing˙page** Which landing page the user see.

**converted** Whether the user converted or subscribed to

old˙page / new˙page Factor 0 / 1 Factor

the course.

## Exploratory Analysis

**1.1**

Table 2. Old vs New Page for Control and Treatment Groups

*old page new page control* 145274 1928

*treatment* 1965 145311

Let’s check the frequency of each control group and the pre- sented landing page. The user who belongs to the con- trol group should get the old destination page, while the processing group should get the new page. It seems that some users are directed to the wrong landing page, so we will filter out the data that gets the correct landing page only.

Figure 1 describes the conversion rates for the Old Page and New Page. Since the conversion rates for both pages are approximately twelve percent, we cannot deduce from the observed data which landing page has a higher conversion rate. We therefore need to do a Bayesian AB test to further investigate.

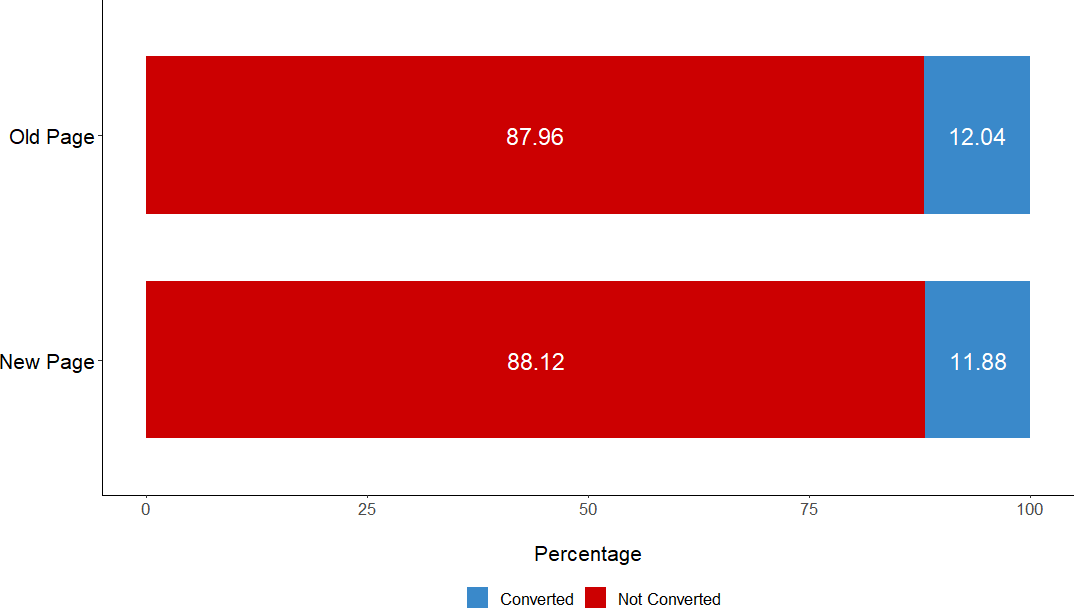


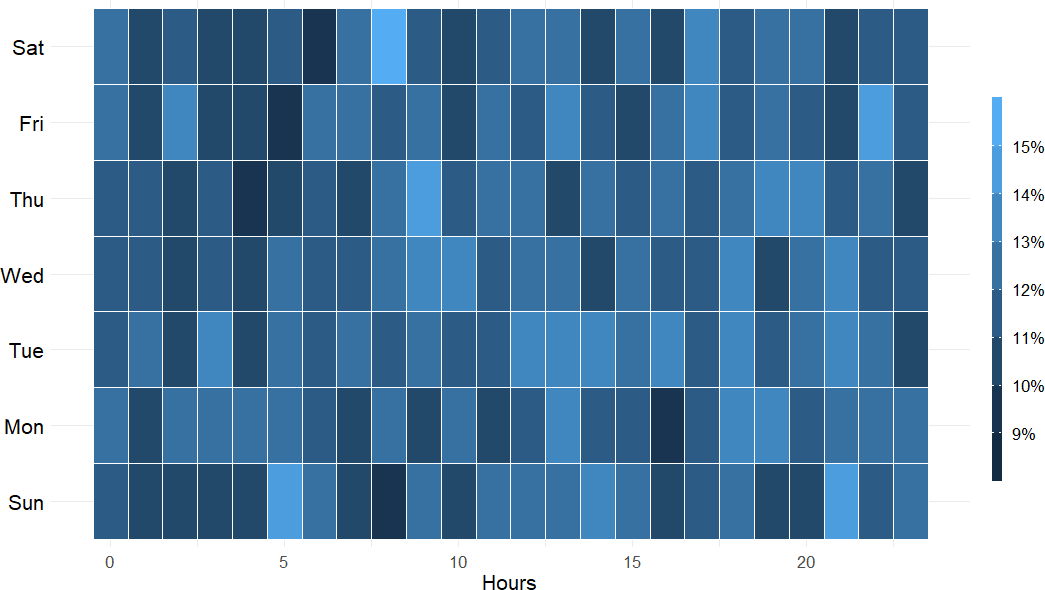
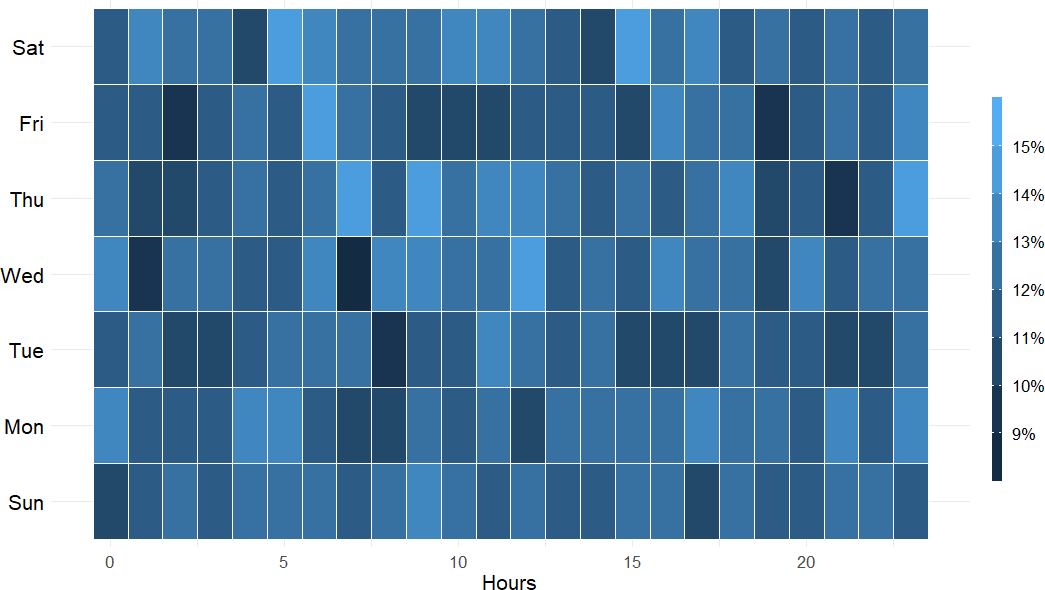
Figure 1. Conversion Rate for Old vs. New Landing Page

To better understand if the day of the week or the hour of the day has an impact on the conversion rate, we need to visualize it. A heatmap which displays the conversions rates for Old and New pages based on days of the week and hours of the day is displayed in Figure 2. For the New Page, the heatmap shows a decrease in conversion rates in the early hours of the weekend. Otherwise, the heatmaps show that there is no significant change in conversion rates for both pages based on days and hours.

## AB Testing

**1.2**

Instead of calculating p-values, a Bayesian AB test was conducted. AB testing is used to compare two versions of something to figure out which performs better. AB testing is commonly used in the field of information technology to test features related to user experience, website landing pages, marketing campaings, etc. In this project, we are comparing two different website landing pages (Old Page and New Page) to see which one has a higher conversion rate. To continue with the AB test, a prior needs to be created. We define a prior of Beta(alpha=2, beta=20) whose density curve is given in Figure 3.

(a) (b)

Figure 2. Heat Map - Conversion Rate For (a) New Page (b) Old Page

### Prior Density Curve

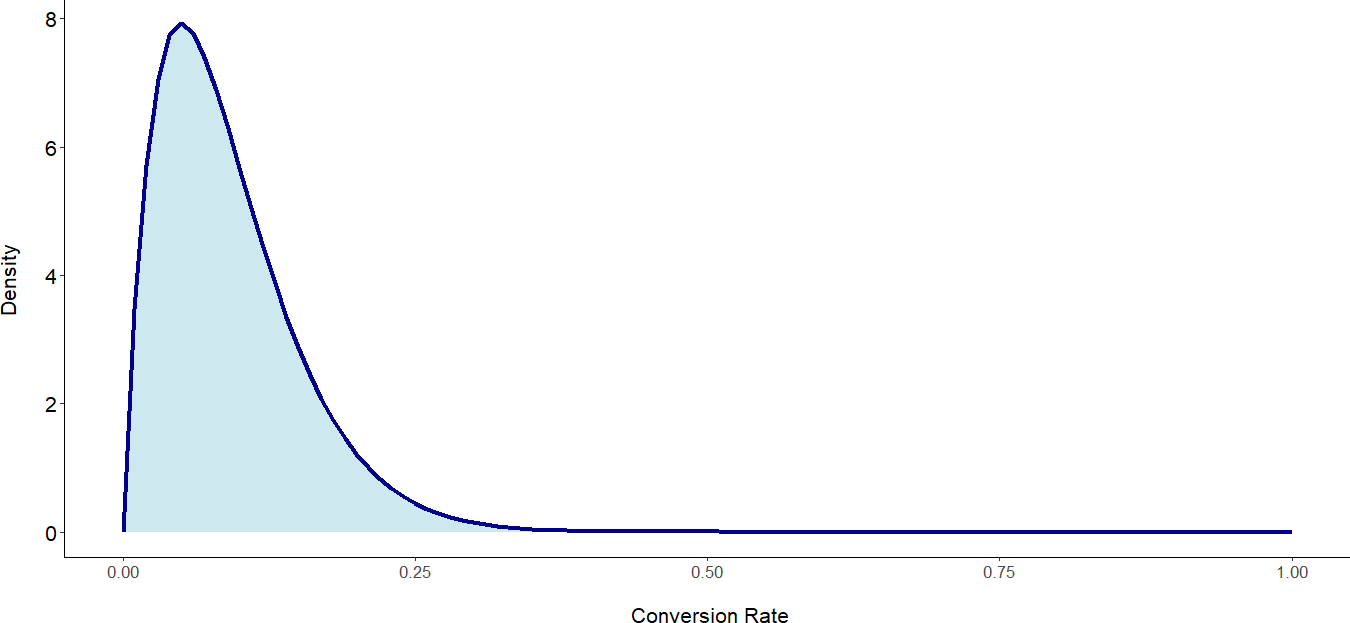


Figure 3. Beta Distribution with Parameters: alpha = 2, beta = 20

### Observed Data

Table 3. Conversion Rate Based on the Observed Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **group** | **landing˙page** | **total˙visit** | **converted** | **not˙converted** | **conversion˙rate** |
| control | Old Page | 145274 | 17489 | 127785 | 0.1203863 |
| treatment | New Page | 145311 | 17264 | 128047 | 0.1188072 |

Table 3 describes the conversion rates for Old Page and New Page based on the experimental data.

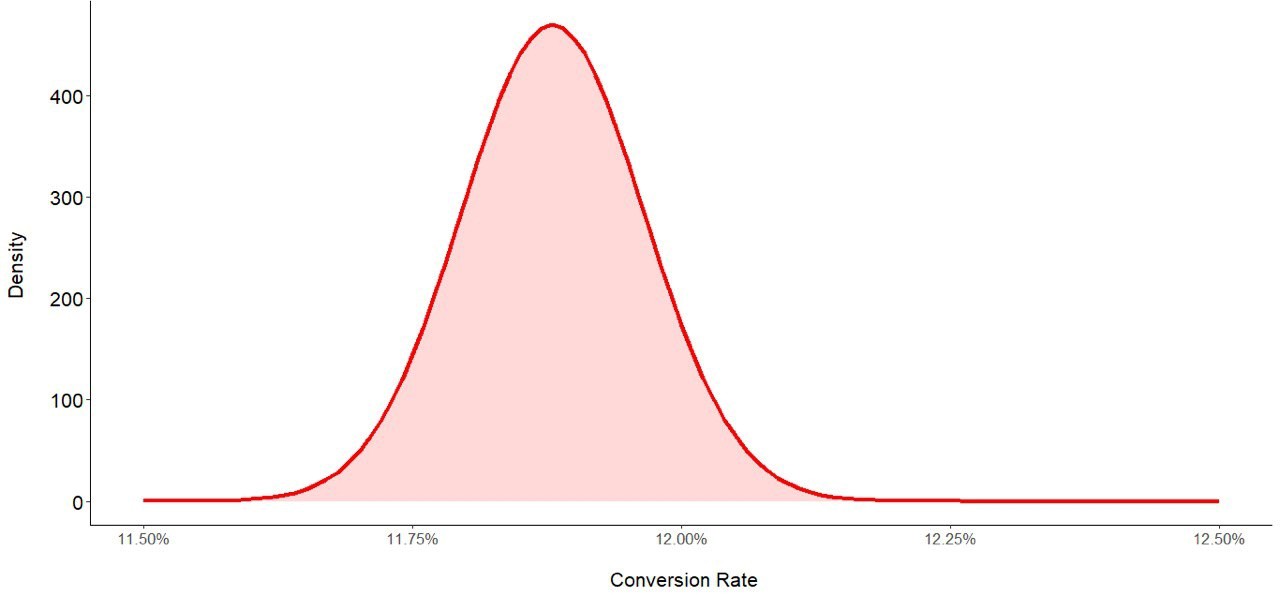
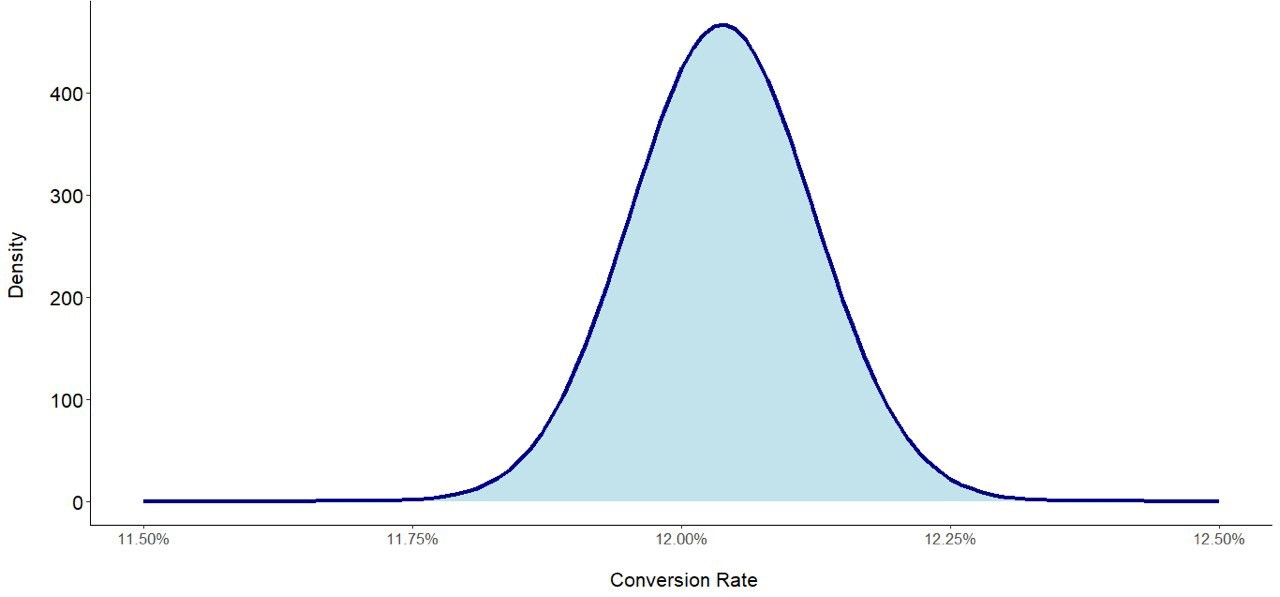
### Posterior Parameter Estimation

The posterior probability curve for each group (Old Page and New Page) is determined by incorpo- rating the information from the prior and the observed data using the beta distribution.

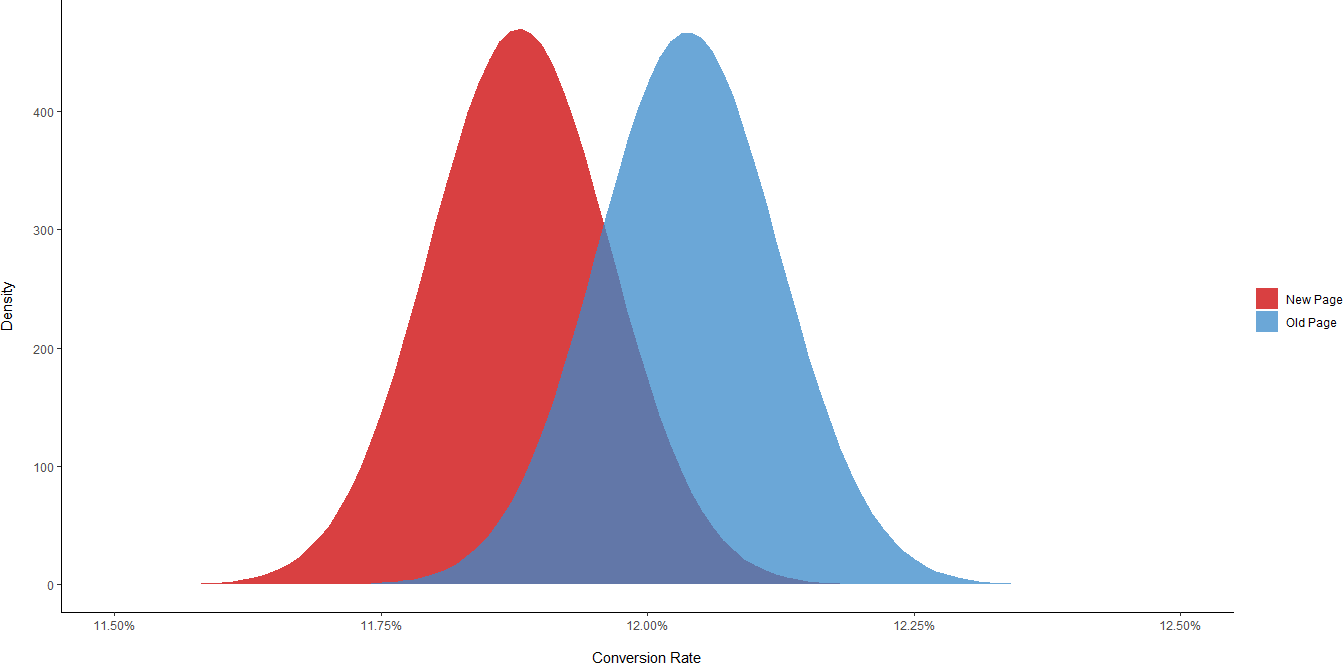


Figure 4 visualizes the density curves for the posterior probabilities of conversion rate for the Old Page and New Page. The posterior distribution is supposed to represent all possible values of conversion rate for each page. It can be clearly seen that the Old Page has a higher conversion rate than the New Page when plotted on the same axes. Based on the density curves, the conversion rate for the Old Page lies between 12.0 percent to 12.25 percent whereas the conversion rate for the New Page lies between 11.75 percent and 12.0 percent.

### Posterior Density Curves



(a) (b)



(c)

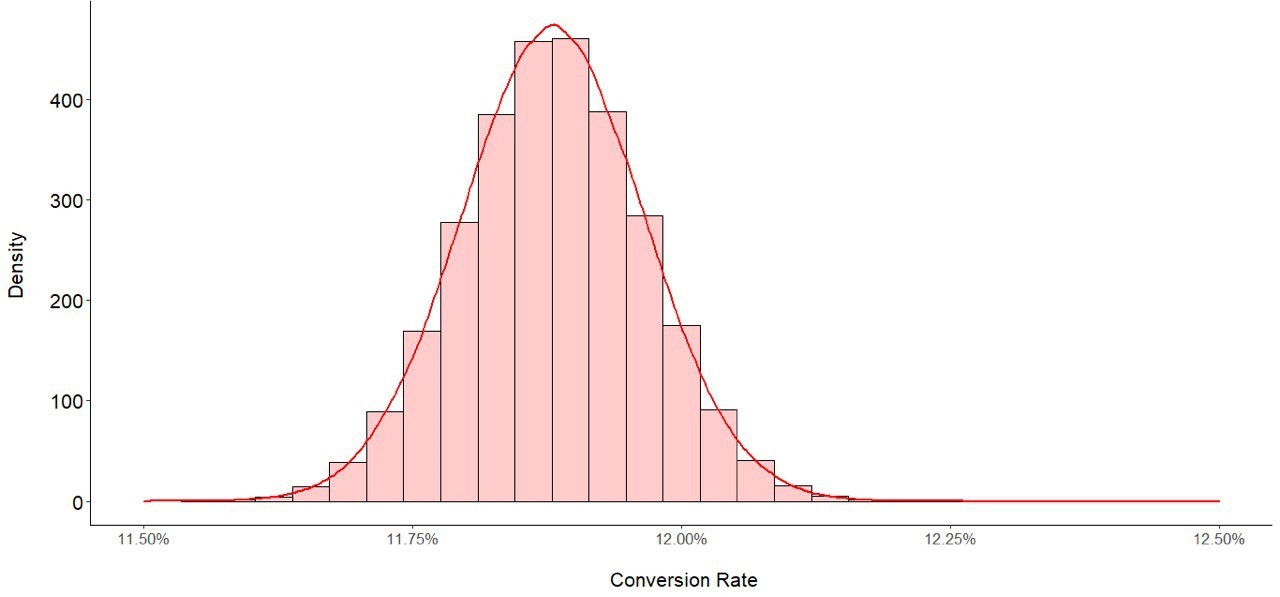
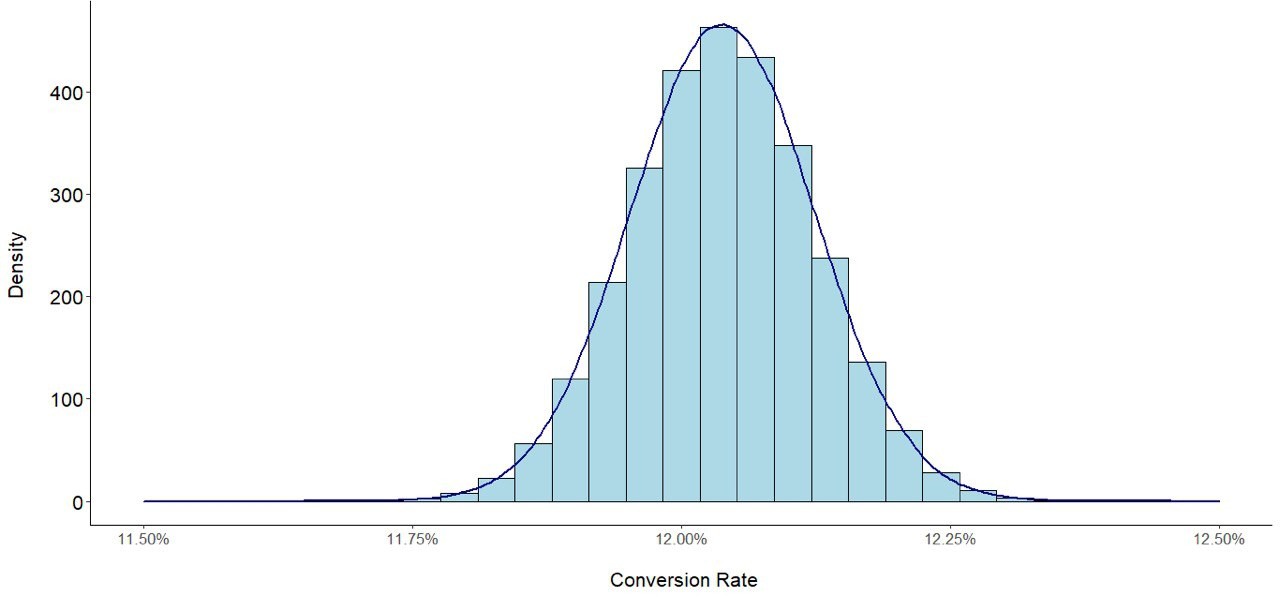
Figure 4. Posterior Probabilities for Conversion Rates of (a) Old Page (b) New Page (c) Old Page and New Page

The data from the posterior probability curves suggests that the Old Page has a higher conversion rate. Nevertheless, we know that the true conversion rate comprises of a range of possible values and that there is some overlap between the true conversion rates for the Old Page and New Page. It cannot be said with certainty that the Old Page has a higher conversion rate. There is a possibility that by chance, the New Page responses were bad in the experiment and it is in fact the page with the better conversion rate.

To answer with certainty, we must simulate the posterior probabilities for each page type using a Monte Carlo simulation with a large number of trials. After random sampling from the two distributions (Old Page and New Page) and each sample being chosen based on its probability in the distribution, it is expected that samples in a high probability region appear more frequently.

After Monte Carlo simulation with 1, 000, 000 trials, the distribution of the posterior probabilities for the Old Page and the New Page is shown in Figure 5. The density curves for each page type represent the estimated conversion rate parameter. For the Old Page, the median for the conversion rate is approximately 12.04 percent. For the New Page, the median for the conversion rate is approximately 11.88 percent. Now, it can be said with more certainty that the Old Page has a higher conversion rate. There may not be a need to change to a new website landing page based on our results.

### Posterior Probabilities after Monte Carlo Simulation with 1, 000, 000 trials



(a) (b)

Figure 5. Posterior Probabilities Distribution for (a) Old Page (b) New Page (after Monte Carlo Simulation with 1M Trials)

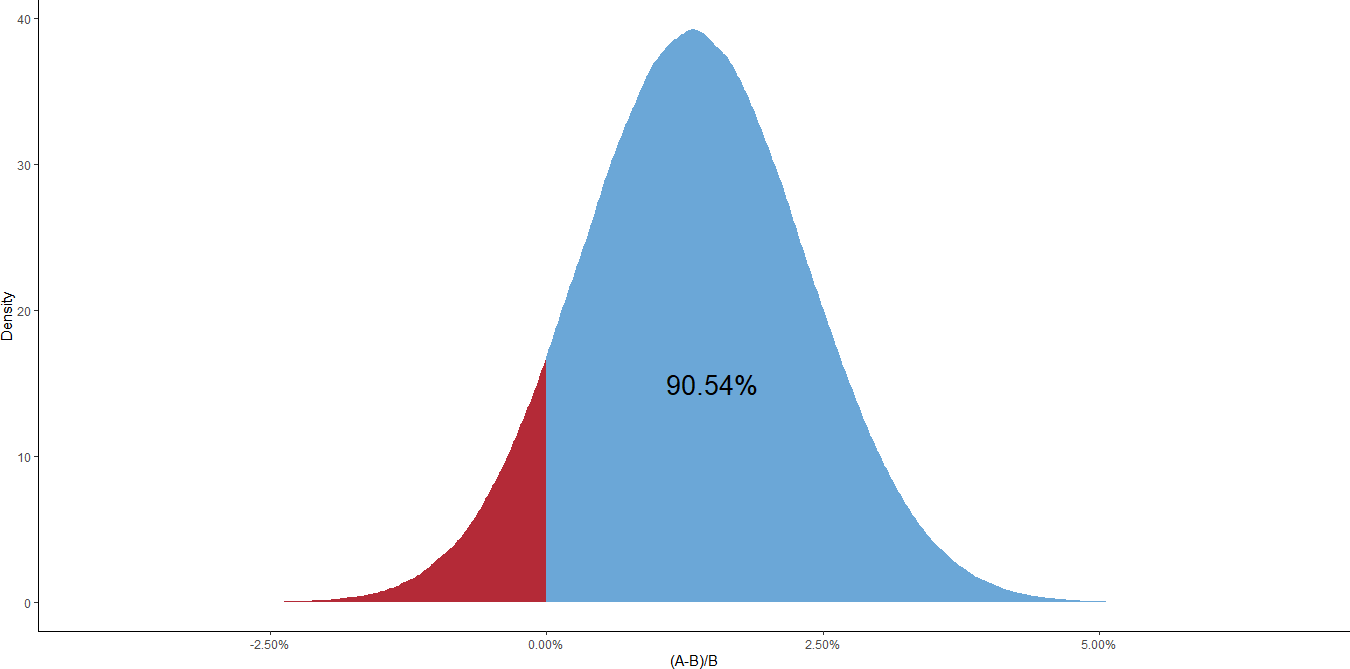
## Convergence of Posterior Probabilities

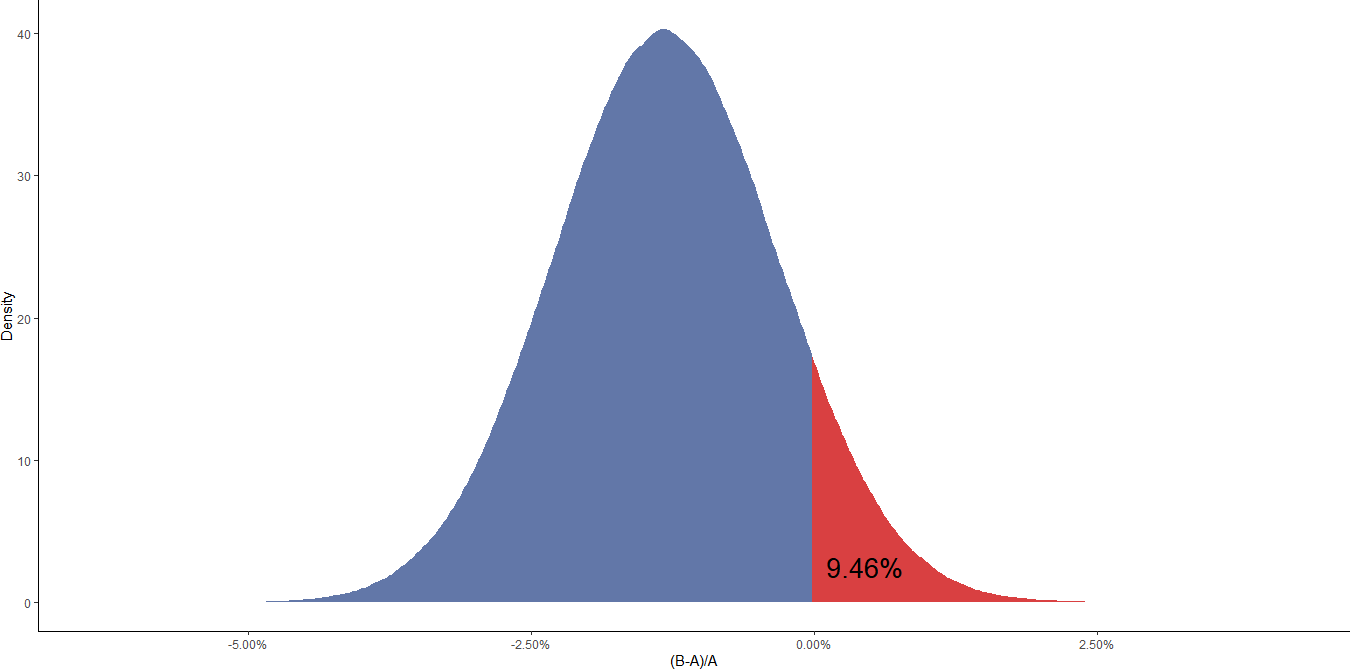
**1.3**

To determine the percentage that the Old Page has a higher conversion rate, we can add the total number of times the Old Page had a higher conversion rate than the New Page in our Monte Carlo experiment and divide it by the total number of trials (1 million). The same method can be used to calculate the percentage that the New Page has a higher conversion rate by reversing the page types. This value is similar to the p-value obtained by frequentist methods, although it is much more intuitive and interpretable.

Figure 6 (a) shows that in 90.54 percent of the one million trials, the Old Page has a better conversion rate. The x-axis depicts the percentage of how much the Old Page is better than the New Page in terms of conversion rates. Based on our prior belief and the observed data, we can conclude that the Old Page has the better conversion rate.

Figure 6 (b) shows that in 9.46 percent of the one million trials, the New Page has a higher conversion rate. **Therefore the posterior probability that the New Page has a higher conversion rate is 9.46 percent.**



(a)

(b)

Figure 6. Posterior Probabilities that (a) Old Page (b) New Page have Higher Conversion Rate

# PREDICTING LENGTH OF STAY IN HOSPITAL

**2**

MIMIC (Medical Information Mart for Intensive Care) is a large, freely-available database that comprising those health-data from patients admitted to the critical care units of the Beth Israel Deaconess Medical Center. It has three versions. MIMIC-II from 2001-2008. MIMIC-III from 2001-2012. MIMIC-IV from 2008-2019, Those data were collected from two different clinical system: CareVue and Metavison.

## Exploratory Analysis

**2.1**

In order to better explore the data, we did several data visualization to know the dataset better.

We do some research on the insurance in Figure 7. Since most of the patients have insurance to cover part of their fees in the hospital otherwise it’s so expensive and unaffordable to common families. In order to compare the length that patients with different types of insurances spent in hospital, we got the average of staying days which is 10.11. The self-pay option definitely is the shortest stay with 6.41, since the more they stay the more they pay and it’s so expensive.

All the rest don’t make any significant difference. One thing that is worth declaring is that, Medicaid is a federal program for those low income people of every age, those people usually pay no part costs for that; and Medicare is the insurance program that for those 65 age or younger and those disabled, and small premium are required monthly for non-hospital coverage. Till now it’s not hard to see that why those patients with Medicaid staying longer than the others, since they

Table 4. Data Dictionary for the Mimic 3D Dataset

**Variable NAME DEFINITION CODES DATA TYPE**

**Age** Age of the patient at the time of hospitalization

**LOSdays** Number of days that patients spent in the ICU

from 0 to 89 Non-negative integer

From 0 to 294.63 Decimal

**Admit˙type** how patients were admitted to the ICU

**Insurance** The type of insurance use by patients

**marital˙status** Marital status of the patient

Elective, Emergency, Newborn and Urgent

Government, Private, Medicaid, Medicare Self Pay

Divorced, Married, Single, Na, Unknown, Life Partner, Separated And Widowed

Factor Factor

Factor

**NumTransfers** Number of time patients transferred from hospitals

to hospitals

**NumDiagonsis** Number of Diagnosis by doctors

**LOSgroupNum** Categories of Length of

From 0 to 125 Decimal

From 0 to 450 Decimal

0, 1,2,3 Factor

Stay in the ICU (LOS days)

are low-income and even some parts of them are homeless. They don’t have enough nutrition to support the function of bodies, consequently staying longer time in hospital than other people.

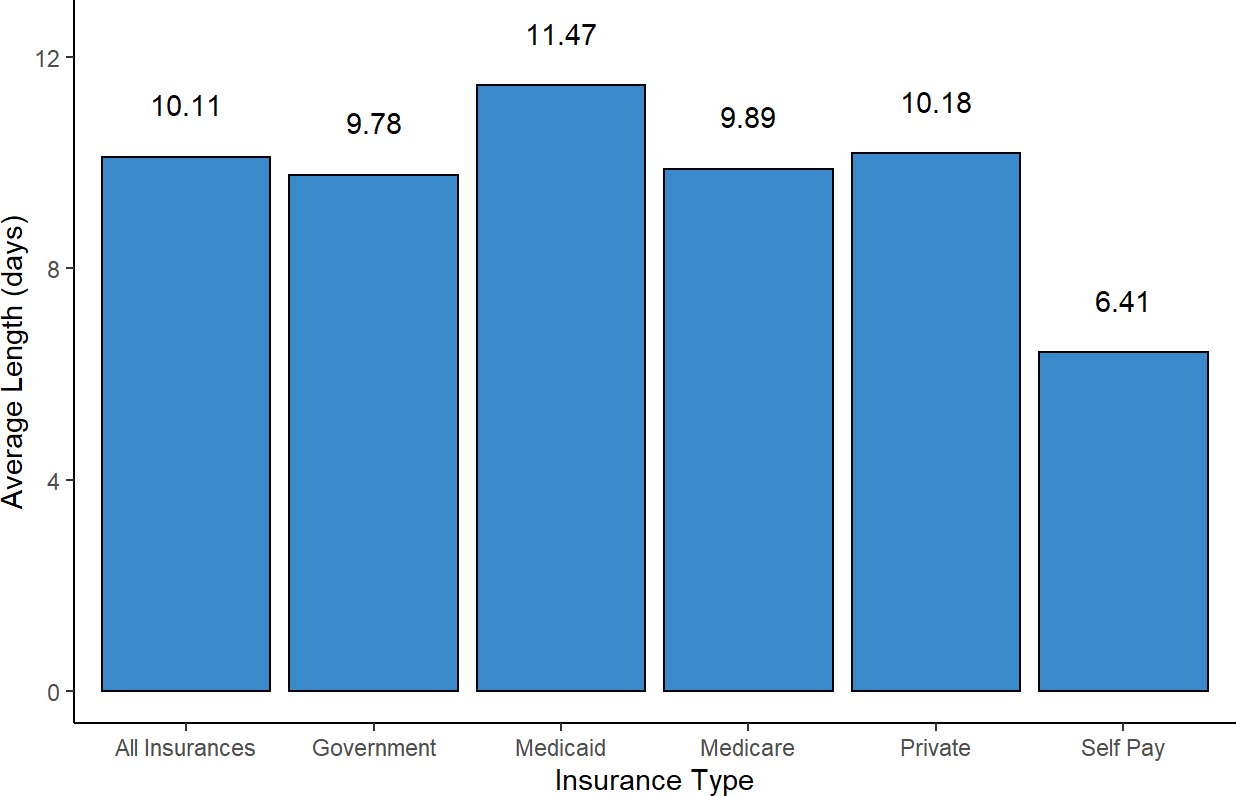


Figure 7. Different Types of Insurance Average vs Total

The admit type can also reflect some information. So we again compare each type’s average to the total average to see the difference in the Figure 8. First is the URGENT, since in usually cases, urgent patients need more time to recover because they may experience fatal and really urgent diseases that must stay in hospital longer than others to make sure they are good. For those new born babies, their immune systems haven’t been developed yet, so they also need more time to leave the hospital. Elective patients may transfer from other hospital or referred by others. As a result, their illness may not be serious or already been cured, so their length of stay is the shortest with average 8.86.

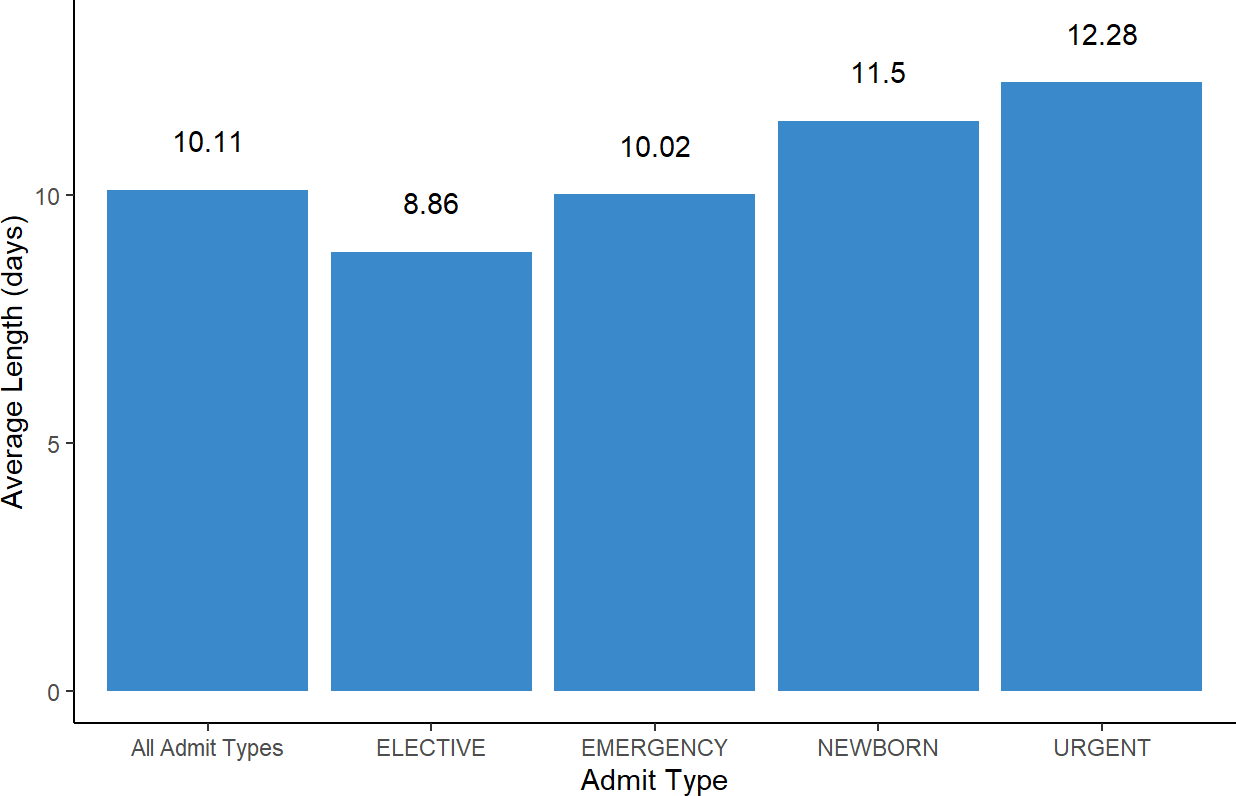


Figure 8. Admit Type vs Length of Stay in Hospital

We try to figure that if there’s some relationship between the patients’ ages and the length of stay in hospital. We first notice that in the Figure 9 . For the Age 0 that are new born babies, the density of new born babies is much higher than other age intervals’ patients. Because most of the new born babies who need to in the hospital usually take a long time for nurses to take care since they are more immunocompromised than adults. Then we turn our sight to above elder patients, we can clearly see that the density of points increase dramatically with the age growing to around 50 years old and reaching a peak.Then it decreases a little after the 50 but still more than those below 25. We understand that as when people reach their menopause, the frequencies of several fatal diseases increase as modern people don’t usually keep their bodies well, especially for people in those ages, busy at work and families make them more possible to get sick. For those elderly, they do a lot more on health care than those people in menopause. However, aging is not reversible. So comparing to youth, they definitely stay more days in the hospital.

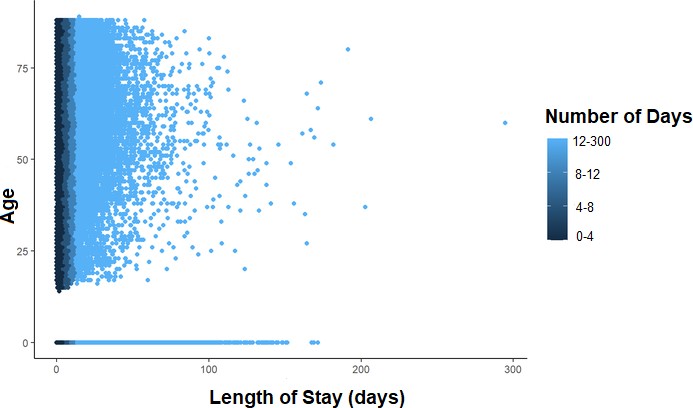


Figure 9. Scatter Plot - Length of Stay in Days vs Age

Marital status will also impact the length of stay in hospital (Figure 10). But only marital status seems far from what we need. So we combine it with insurance type to show more. As we can see in the Self Pay, most of them are in [0, 4), meaning that whatever the marital status is, most of the patients stay less than 4 days. Another worth discussing is Medicaid. Most of them are longer than 12 days, whatever marital status is, since it’s for low-income people and patients almost don’t need to pay at all. Insurance coverage can be deduced from this heat map as well. Medicare has the least patients who stay in hospital longer than 12 days except Self Pay. Combing with the knowledge that it’s for 65 years and below, it confirms our thoughts. Government and Private insurance come the second as three blocks of each larger than 12 days.

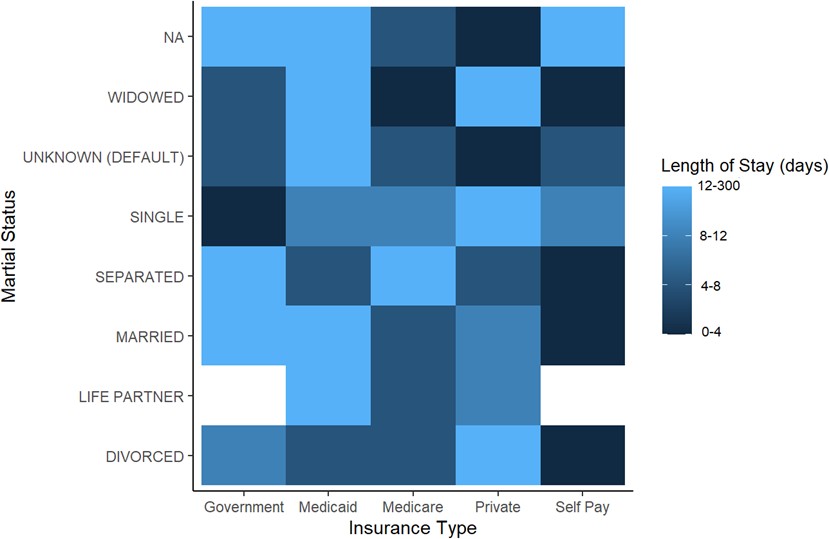


Figure 10. Heat Map - Insurance Type vs Martial Status in Length(Days)

## Bayesian Linear Regression Analysis

**2.2**

Bayesian linear regression analysis differs from frequentist linear regression in that it is focused on probability distributions and not point estimates. The response variables is not estimated to be a single value but is believed to be one of the possible values from a probability distribution. For a response variable sampled from a normal distribution, the Bayesian Linear Regression Model is:



The main focus of Bayesian linear regression is to find the posterior distribution for the model parameters and not a single optimal value of them. Both the response variable and the model parameters are derived from probability distributions.

1. Bayesian Linear Regression Model

We form our model by selecting variables that are correlated to the length of stay (LOS days) to make this model more reasonable. Since there are too many variables in the dataset mimic3d, we remove variables that can be judged by common sense and our calculation. We choose variable insurance as an example in Figure 11. Then we make a model with only insurance to the length of stay ( LOS days). It is obvious that some of their 95 percent confidence intervals do cross the 0, meaning that if we run this experiment again, there is a good chance of finding no correlation with the length of stay (LOS days) for the insurance variable. So we leave those variables whose confidence interval contain 0, and are left with these variables: NumCallouts, age, admit˙type , NumDiagnosis, NumProcs, NumCPTevents, NumInput, NumLabs, NumMicroLabs, NumNotes, NumOutput, NumRx, NumProcEvents, NumTransfers, NumChartEvents and TotalNumInteract.

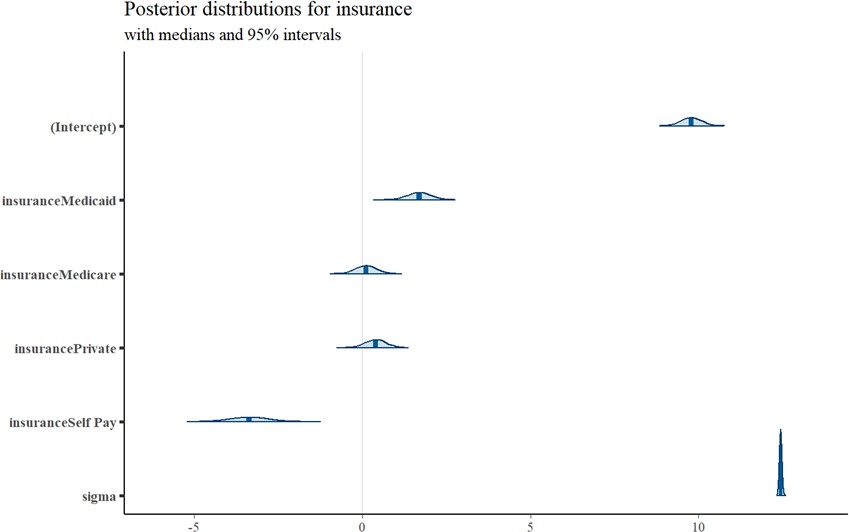


Figure 11. Posterior Distribution for Insurance

### Select, Fit and Predict the Regression Model using *rstanarm* package

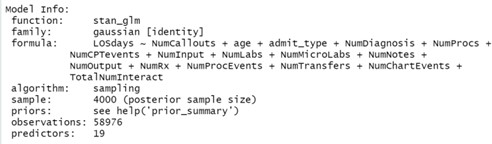


Figure 12. Model Information

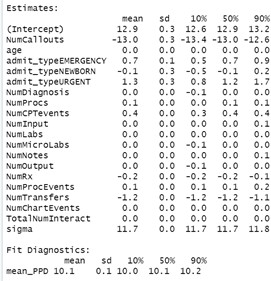


Figure 13. Estimates and Fit Diagnostics

Then we combine all those variables and form a linear regression model to see how they correlate to the length of stay (LOS days). After it we apply this model to the patients.xlsx and make prediction.

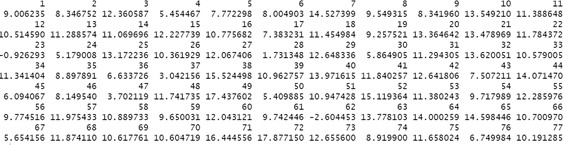


Figure 14. Predicted Values

### What is the Probability of Staying Longer than 2 days in the Hospital?

The pnorm function represents the cumulative probability that the length of stay will shorter than 2 days. To get the probability that the length of stay in the hospital will be longer than 2 days, we subtract the cumulative probability from one.

1 *− pnorm*(2*, mean*(*Y* )*, sd*(*Y* )) = 1 *−* 0*.*01327981 = **0.98672019**

where Y represents the prediction of length of days spent in hospital (LOS days) using our linear regression model.

The probability of staying longer than 2 days in the hospital is 98.67 percent which means that almost all patients will stay longer than 2 days in the hospital. This could be due to a number of reasons, one of which might be that the data in the ”patients” dataset is of very bad quality, with a lot of missing and extreme values when compared to the ”mimic3d” dataset. So the ”patients” dataset might be very different from the ”mimic3d” dataset which is why our probability prediction is so extreme.

# CONCLUSION

In conclusion, Bayesian analysis offers a natural and structured way of combining prior beliefs with observed data to create a robust decision-making framework. This report introduced two applications: Bayesian AB testing and Bayesian linear regression analysis. Both applications differed from frequentist methods in that they were focused on probability distributions rather than point estimates. Compared to a p-value which is often misinterpreted, the Bayesian framework allows us to find the probability directly which can be understood easily and help drive action. The ability of Bayesian analysis to incorporate prior beliefs and allow more sophisticated analyses is resulting in an increase of use of Bayesian methods in the fields of medicine, law, business, and philosophy where unobserved data is often needed for model creation.

# REFERENCES

### List of References

Introduction to bayesian linear regression - towards data science. (n.d.). Retrieved October 29, 2022, from https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

Boemer, R. (2021, October 25). 5 concrete benefits of Bayesian statistics. Medium. Retrieved October 29, 2022, from https://medium.com/p/9acae2a8ae0d

MIMIC-III documentation. (n.d.). MIMIC. Retrieved October 29, 2022, from https://mimic.mit.edu/docs/iii/

RPubs - Bayesian Statistics and A/B Testing. (2021, January 28). https://rpubs.com/Argaadya/bayesian˙ab

User Dikran Marsupial. (n.d.). Cross Validated. Retrieved October 29, 2022, from https://stats.stackexchange.com/users/887/dikran-marsupial

With R, M. (2020, April 25). Bayesian linear regression. R-bloggers. h[ttps://www.r-bloggers.com/2020/04/ba](http://www.r-bloggers.com/2020/04/bayesian-linear-regression/amp/)y[esian-linear-regression/amp/](http://www.r-bloggers.com/2020/04/bayesian-linear-regression/amp/)