VACCINE REPORTING IN CALIFORNIA

A statistical Analysis of Vaccine Reporting Data from Kindergartens in California

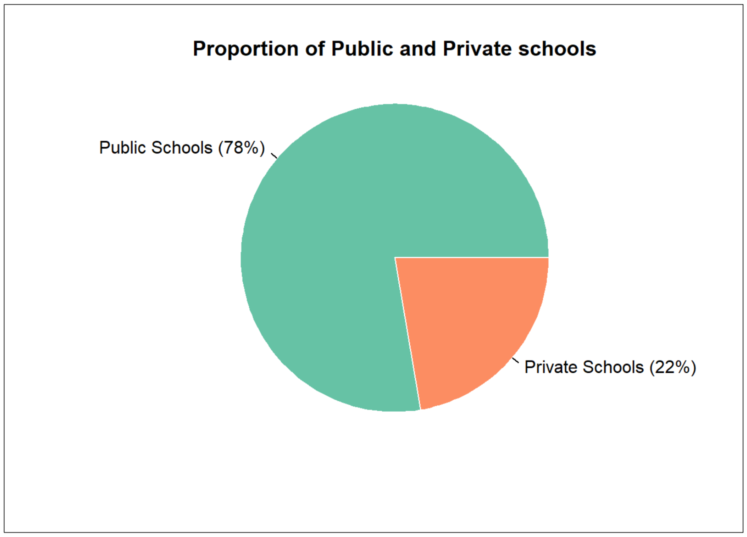
STEPHEN OMONDI – SYRACUSE UNIVERSITY

IST 772 - Quantitative Reasoning in Data Science

CONTENTS

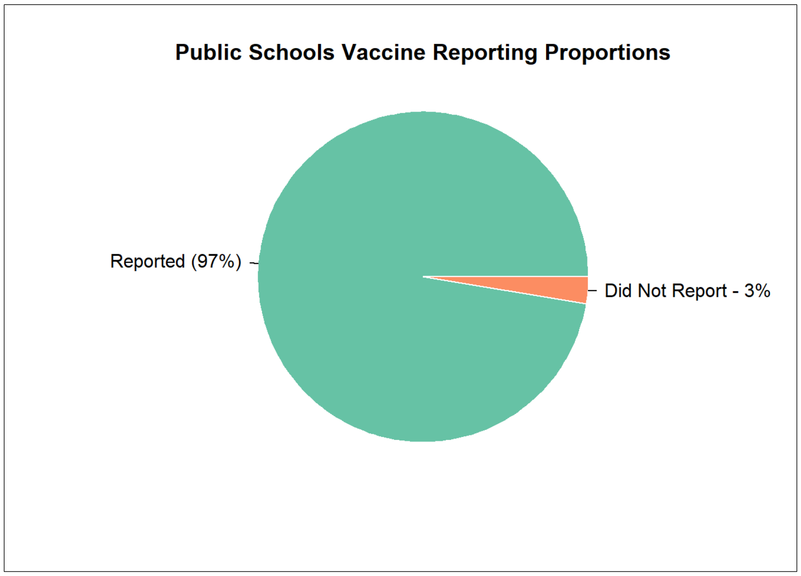
INTRODUCTORY/DESCRIPTIVE REPORTS

Overall, there were a total of 7381 Kindergarten schools under the survey. 5732 of these were Public schools which accounted for 78% of the total while 1649 were private schools accounting for 22% of the total.



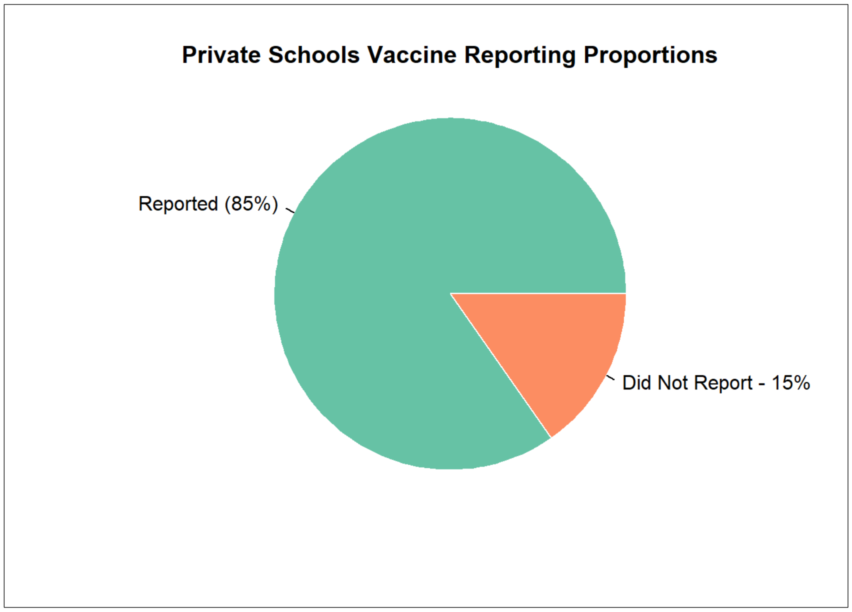
What proportion of public schools reported vaccination data?

5584 public schools equivalent to 97% of all public schools reported vaccines while 148 private schools equivalent to 3% did not. However, the public school proportion reporting vaccines accounts for 76% of all kindergarten schools within the state, while the proportion that did not report vaccines accounts for 2% of all kindergarten schools within the state.



What proportion of private schools reported vaccination data?

1397 private schools equivalent to 85% of all private schools reported vaccines while 252 private schools equivalent to 15% of private schools did not. The private school proportion reporting vaccines accounts for 19% of all kindergarten schools within the state, while the proportion that did not report vaccines accounts for 3% of all kindergarten schools within the state.



Have U.S. vaccinations rates been stable over time?

According to the survey data between 1980 – 2017, emerging trends in US Vaccinations can be summarized as follows.

- **Polio third dose (Pol3)**
  There was an upward trend in U.S Polio third dose vaccines administered for the period covered. This upward trend appears to have been nearing its plateau.

- **Hepatitis B, Birth Dose (HepB_BD)**
  There was a strong upward trend in Hepatitis B, Birth Dose vaccines from the first recorded time in the year 2000.

- **First dose of Diphtheria/Pertussis/Tetanus vaccine (DTP1)**
  There was a sharp downtrend in DTP1 vaccine for the recorded time.

- **Influenza third dose (HiB3)**
  There was a sharp downtrend in HiB3 vaccines for the recorded period.

- **Measles first dose (MCV1)**

  Overall, the MCV1 vaccines appear to have plateaued.

These summaries were arrived at after decomposing the available time series data into its constituent parts in order to examine trends, seasonality, cyclicity and irregularities – if any. *Figure 1* below shows the vaccination time-series data before decomposition while *Figure 2* shows the plotted data after decomposition.

```
plot((usVaccines))
```

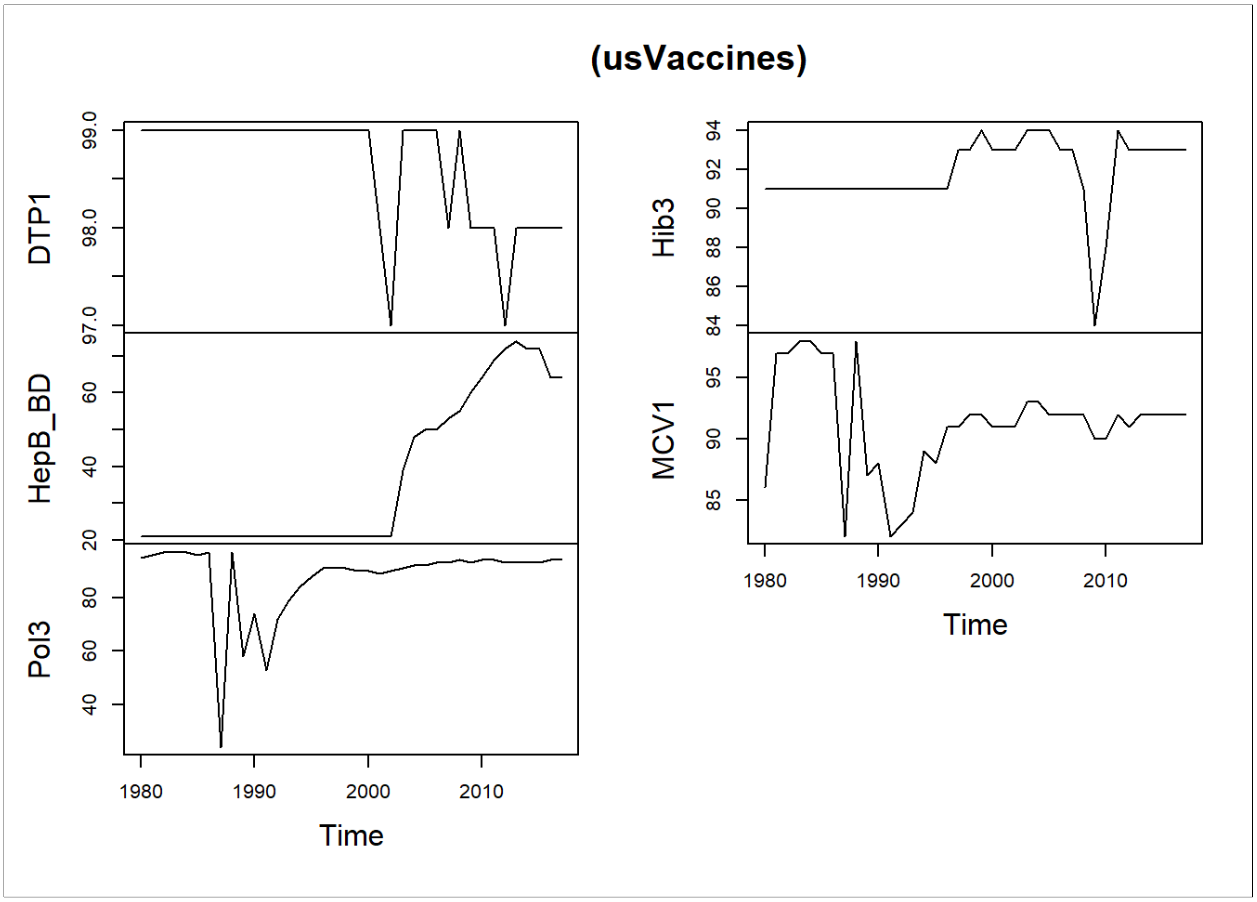


Figure 1: US Vaccination time-series data before decomposition

```
# convert the period into meaningful time i.e. every 12 calendar months
freq <- 12
decVac <- decompose(ts(usVaccines, frequency=freq))
plot(decVac$trend)
```

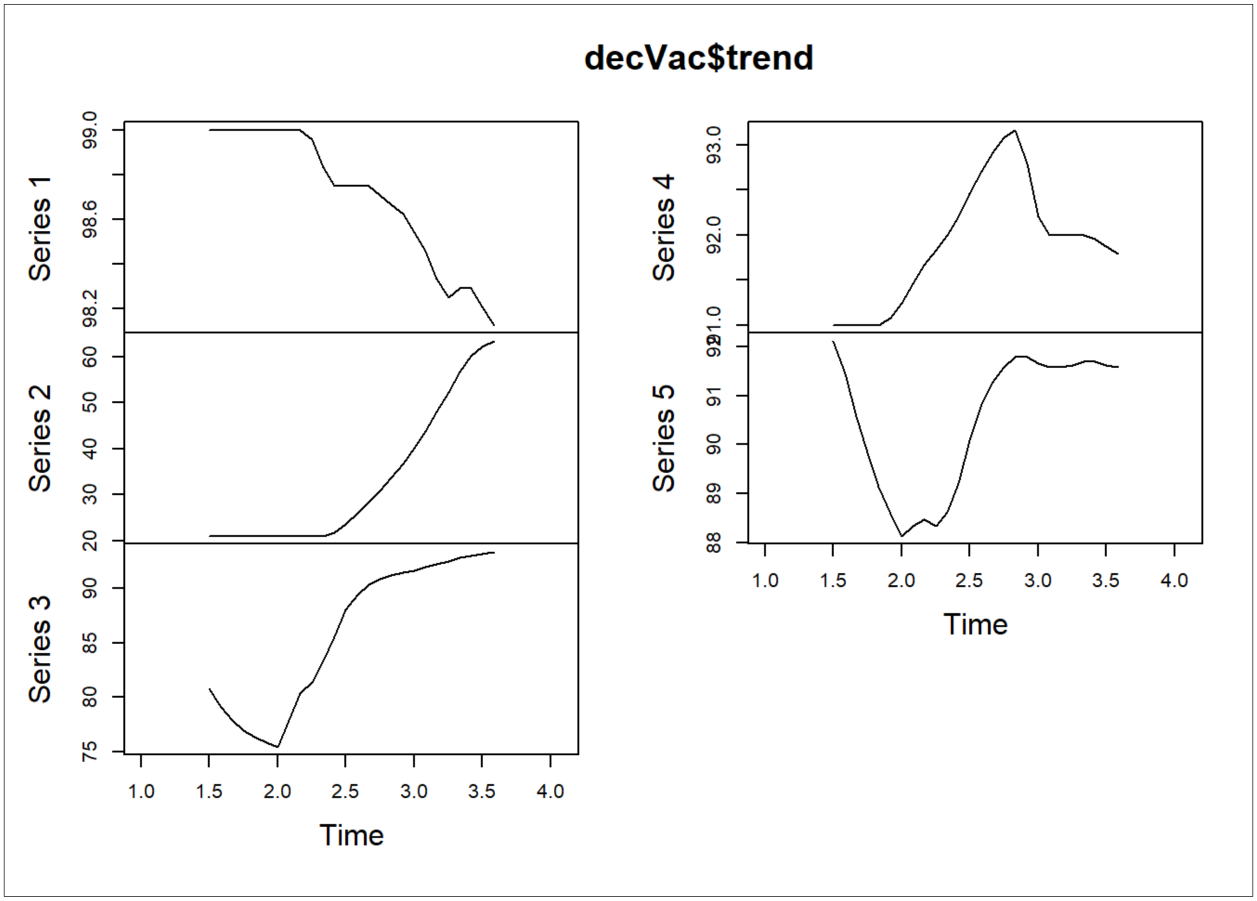


Figure 2: US Vaccination time-series data after decomposition

Are there any notable patterns in U.S. vaccinations rates over time?

Even though there were observed trends as mentioned above, the US vaccination data showed very little seasonality. In fact, the mean seasonality was a mere 0.02. *Figure 3* visualizes potential seasonality as shown below:

```
mean(decVac$seasonal, na.rm = TRUE)
## [1] 0.02369052
plot(decVac$seasonal)
```

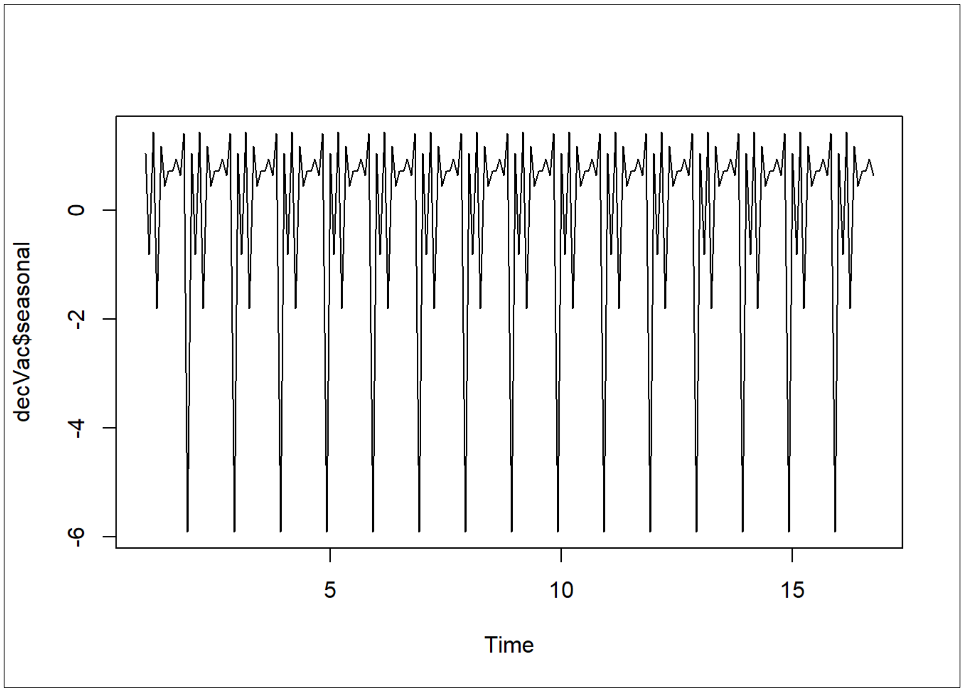


Figure 3: Mean Seasonality is very small (~0.2)

The trend component inherent in the time-series was removed to limit the potential for spurious correlations among different vaccines. This was done through a differencing approach that establishes stationarity. Figure 4 below shows the vaccines after differencing.

```
diffSet <- diff((usVaccines))

#head(diffSet)

plot(diffSet)
```

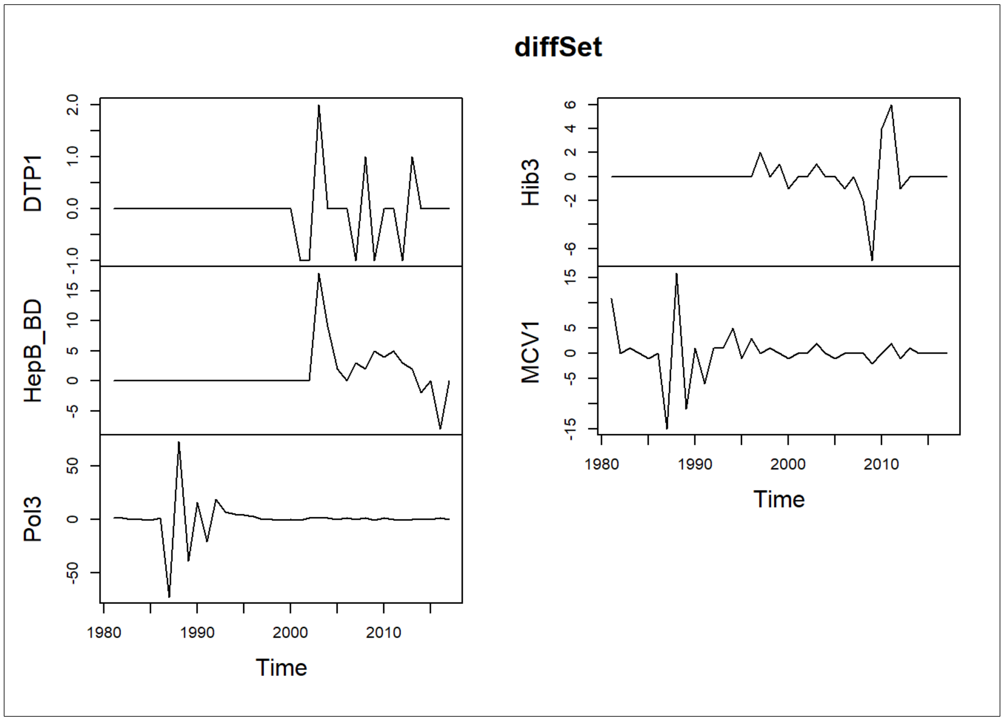


Figure 4: US Vaccinations after Differencing to establish stationarity

The charts show differences in variability at certain times (heteroscedasticity) which points to spikes and dips in vaccinations at those times. For example, there was increased variability between 1985 and approximately 1994 in the administration of the third dose of polio vaccine (Pol3) which is very similar to the spikes and dips observed in the administration of MCV1 vaccines during the same period. These periods coincide with the recorded goal set for polio elimination in the Americas that started in 1985 which was closely followed by the Global Polio eradication initiative of 1988 that resulted in the polio being declared eliminated from the Americas in 1994 (History of Vaccines, 2020).

Further investigation of the survey data for patterns show strong correlation exists between Measles first dose (MCV1) and Polio third dose (Pol3) vaccine rates. *Figure 5* below shows the positive correlations displayed in blue and negative correlations in red. Color intensity is proportional to the correlation coefficients.

```
library(corrplot)
## corrplot 0.84 loaded
corr <- cor(diffSet)
corrplot(corr, method="number")
```

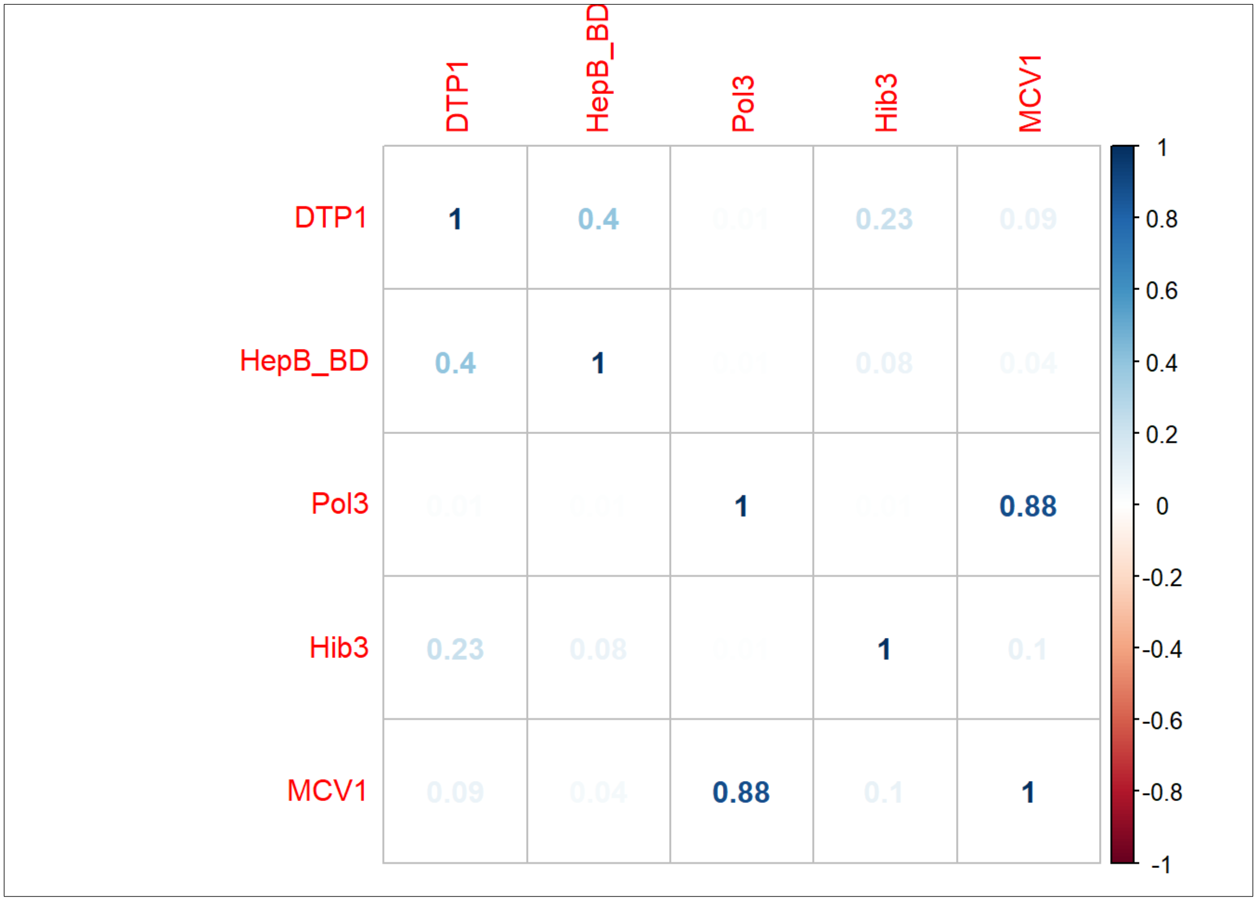


Figure 5: Correlation Matrix of US Vaccines

PUBLIC VS. PRIVATE SCHOOL COMPARISONS

Was there any credible difference in overall reporting proportions between public and private schools?

Overall, there are a total of 7381 Kindergarten schools, in which public schools account for 78% while private schools account for 22%.

97% of all public schools reported vaccines while 3% did not. However, the public school proportion reporting vaccines accounts for 76% of all kindergarten schools, while the proportion that did not report vaccines accounts for 2% of all kindergarten schools.

At the same time, 85% of all private schools reported vaccines while 15% did not. The private school proportion reporting vaccines accounts for 19% of all kindergarten schools, while the proportion that did not report vaccines accounts for 3% of all kindergarten schools.

```
PrivateSchoolBreakdown <- c(countPrivateSchoolsThatReported,countPrivateSchoolsThatDidNotReport)

PublicSchoolBreakdown <- c(countPublicSchoolsThatReported,countPublicSchoolsThatDidNotReport)

# create a matrix of records

AllSchoolsCombined <- matrix(c(PublicSchoolBreakdown,PrivateSchoolBreakdown), ncol = 2, byrow = F
)

# set row names

rowtitles <- c("Reported Vaccines","Did Not Report Vaccines")

rownames(AllSchoolsCombined) <- rowtitles

# set colum names

columntitles <- c("Public Schools", "Private Schools")

colnames(AllSchoolsCombined) <- columntitles

# make a table

AllSchoolsCombined <- as.table(AllSchoolsCombined)

# transpose rows to columns

AllSchoolsCombined <- t(AllSchoolsCombined)

AllSchoolsCombined

##                   Reported Vaccines Did Not Report Vaccines

## Public Schools                 5584                     148

## Private Schools                1397                     252
```

Total Number of Schools in the survey:

```
# grand total of schools

(sumAllSchools <- margin.table(AllSchoolsCombined))

## [1] 7381
```

Public and Private school totals:

```r
# grand total of rows
(sumRows <- margin.table(AllSchoolsCombined,1))
##  Public Schools Private Schools
##            5732            1649
```

Vaccination totals:

```r
# grand total of columns
(sumCols <- margin.table(AllSchoolsCombined,2))
##       Reported Vaccines Did Not Report Vaccines
##                    6981                      400
```

Totals expressed as probabilities (of all schools combined):

```r
# probs expressed as percentages
(schoolProbs <- ((AllSchoolsCombined / sumAllSchools)))
##                  Reported Vaccines Did Not Report Vaccines
## Public Schools          0.75653705              0.02005148
## Private Schools         0.18926975              0.03414172
# probs expressed as percentages
(schoolProbs <- round((AllSchoolsCombined / sumAllSchools) * 100)) #express as %
##                  Reported Vaccines Did Not Report Vaccines
## Public Schools                  76                        2
## Private Schools                 19                        3
(schoolProbs <- ((AllSchoolsCombined / sumAllSchools) * 100))
##                  Reported Vaccines Did Not Report Vaccines
## Public Schools           75.653705                2.005148
## Private Schools          18.926975                3.414172
# overall proportions between public kindergartens and all kindergartens
(round((countPublicSchools/sumAllSchools) * 100,0))
## [1] 78
# overall proportions between private kindergartens and all kindergartens
(round((countPrivateSchools/sumAllSchools) * 100,0))
## [1] 22
# proportions within Private schools
(round((countPrivateSchoolsThatReported/countPrivateSchools) * 100,0))
## [1] 85
(round((countPrivateSchoolsThatDidNotReport/countPrivateSchools) * 100,0))
## [1] 15
# proportions within Public Schools
(round((countPublicSchoolsThatReported/countPublicSchools) * 100,0))
```

```
## [1] 97
(round((countPublicSchoolsThatDidNotReport/countPublicSchools) * 100,0))
## [1] 3
```

Compare overall vaccination rates (allvaccs) between public and private schools. Are there any credible differences?

This report uses a sample data of 698 observations and 13 variables. There are 570 public schools and 128 private schools in the sample data used for this report.

Data Prep and Sampling

100 random vaccine rate sample means are drawn from each school group to ensure equal and repeatable comparison between private and public schools.

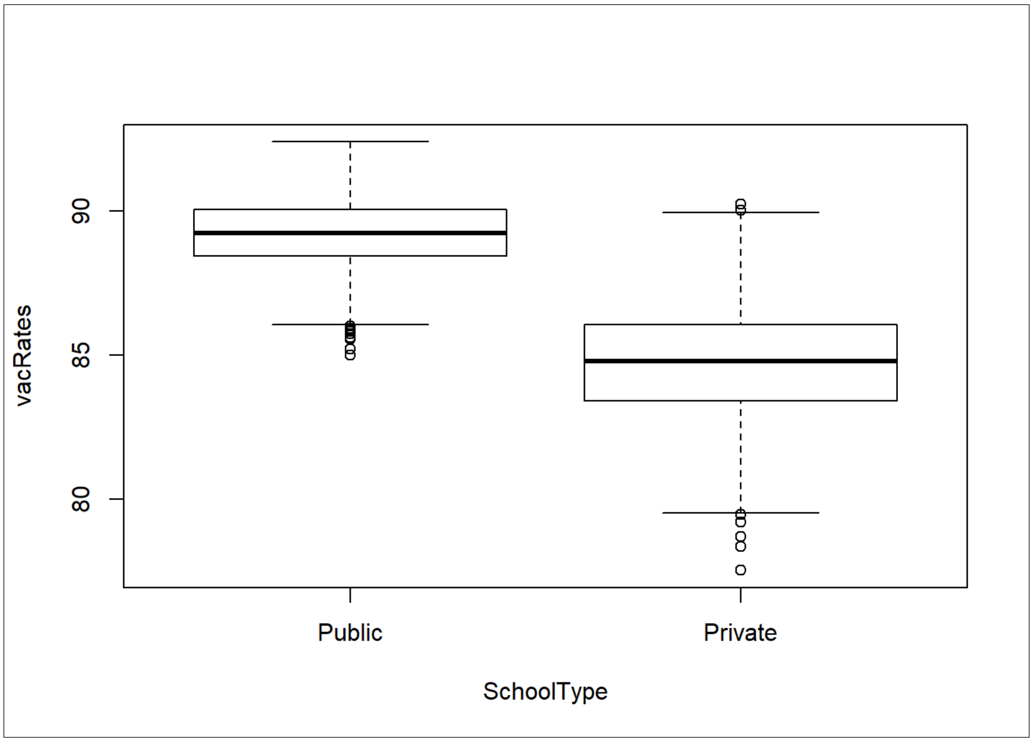
The sample means are replicated 1000 times from each group to create a very large mean samples because as mean samples (size) increases, the means observed are likely to converge to the true population mean and also looks more normally distributed (bell-curved in shape).

```
#str(reportSample)
str(reportSample$allvaccs)
##  num [1:698] 71.4 98.7 86.7 100 98.6 ...
# are there any missing values in allVaccs?
length(complete.cases(reportSample$allvaccs)) #no missing values
## [1] 698
# public schools
PublicSchoolGroup <- reportSample[reportSample$pubpriv=="PUBLIC",]
nrow(PublicSchoolGroup)
## [1] 570
# private schools
PrivateSchoolGroup <- reportSample[reportSample$pubpriv=="PRIVATE",]
nrow(PrivateSchoolGroup)
## [1] 128
set.seed(123) # control the randomization
# obtain 100 random public school vaccine rates sample means, repeat X1000
vacRates <- replicate(1000, mean(sample(PublicSchoolGroup$allvaccs, 100, replace = TRUE)))
SchoolType <- as.factor("Public")
publicDF <- data.frame(vacRates,SchoolType)
# obtain 100 random private school vaccine rates sample means, repeat X1000
```

```
vacRates <- replicate(1000, mean(sample(PrivateSchoolGroup$allvaccs, 100, replace = TRUE)))

SchoolType <- as.factor("Private")

privateDF <- data.frame(vacRates,SchoolType)

# combine both groups into single group of both public and private vaccine rate samples

SchoolVacs <- rbind(publicDF,privateDF)

#View(SchoolVacs)
```

Figure 6 below shows the boxplot of the distributions of private and public sample means after the sampling process above.

```
boxplot(vacRates ~ SchoolType, data = SchoolVacs)
```



.

Figure 6: Boxplot of all vaccines sample means of Private and Public Schools

The box plot shows that the medians of vaccine rates between public school and private schools are very different; public schools have a higher median.

Frequentist approach using ANOVA

```
vacOut <- aov(vacRates ~ SchoolType, data = SchoolVacs)

summary(vacOut)

##               Df Sum Sq Mean Sq F value Pr(>F)

## SchoolType     1  10068   10068    3859 <2e-16 ***

## Residuals   1998   5213       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There were a total of 2000 elements, 1000 each from public and private schools. The degrees of freedom residual elements are 1998 with 1 element left to vary between the two groups during the statistical calculation of ANOVA.
- The Sum Sq value shows the raw initial calculation of variability. This value was 10068 between groups and 5213 within groups.
- The Mean Sq value is the variance (the sum of squares divided by the digress of freedom). This value was 10068 between groups and 3 within groups.
- The F-Value is the ratio of the mean squares between and within groups. This value was 3859
- The Pr(>F) is the probability of a larger F-value of this ratio (3859) within the F-distribution. This value was <2e-16.

For the result to be statistically significant, the F-Value must substantially exceed 1 under the assumption that all the data was sampled from the same underlying population - as is the case here. Also, the Pr(>F) has to be less than the alpha value - in this case, alpha threshold is 0.05.

Frequentist Findings:

- The **null hypothesis** is that there is no difference in sample means of vaccine rates between public and private schools.
- The **alternative hypothesis** is that the sample means of vaccine rates between public and private schools are different.
- Since the Pr(>F) of <2e-16 is less than alpha-value ($p < 0.05$), we reject the null hypothesis and conclude that the test is statistically significant; there is credible evidence that sample means of vaccine rates between public and private schools are different.

Bayesian Approach using Bayes Factor Analysis

The 95% HDI distributions of means between private and public schools was investigated to inform a Bayesian conclusion as shown in the *Figure 7* and *Figure 8* below.

```
vacBayesOut <- anovaBF(vacRates ~ SchoolType, data = SchoolVacs)
mcmcOut <- posterior(vacBayesOut,iterations = 10000)
```

```
par(mfcol=c(1,1))
hist(mcmcOut[,"SchoolType-Private"], main = "95% HDI distribution of means of all vacc
ines in Private School")


# upper and lower bounds of 95% HDI
```

```r
abline(v=quantile(mcmcOut[,"SchoolType-Private"], c(0.025)), col="blue")
abline(v=quantile(mcmcOut[,"SchoolType-Private"], c(0.975)), col="blue")
```

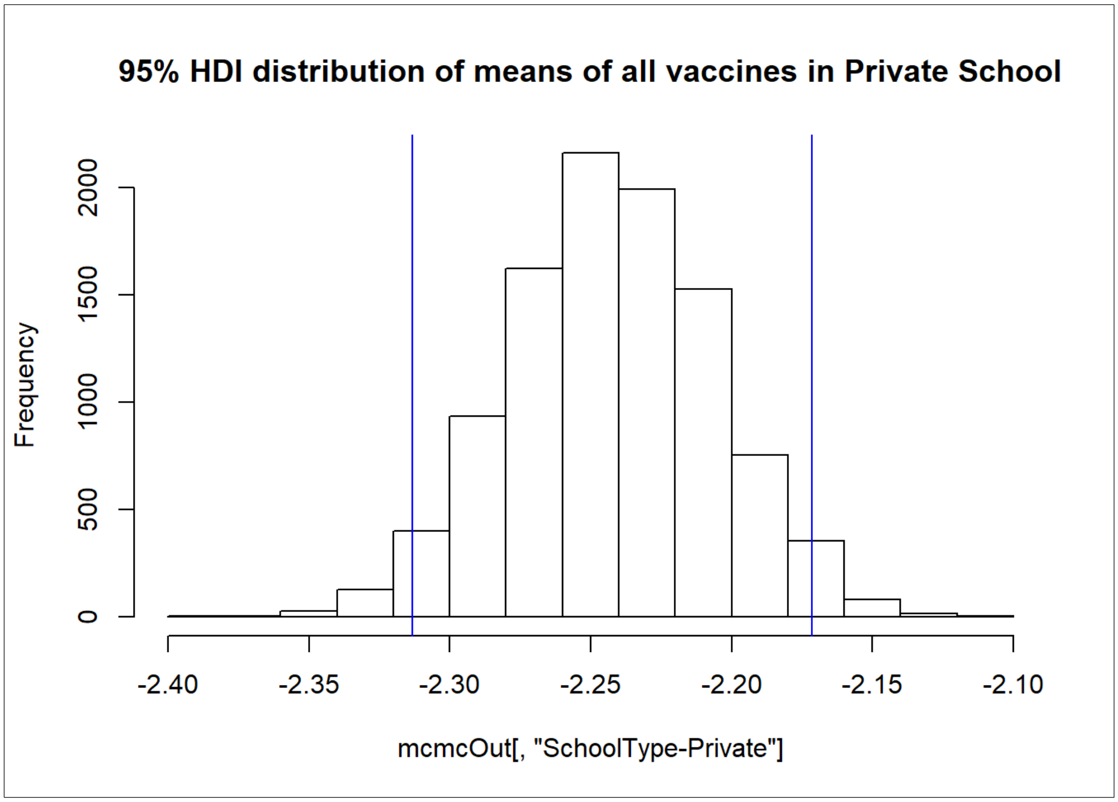


Figure 7: Distribution of Means of All Vaccines in Private Schools

The 95% HDI distribution does not include 0 as shown above which means it deviates from the population mean considerably. The 95% HDI upper bound to lower bounds ranges from -2.2313 to -2.086

```r
par(mfcol=c(1,1))
hist(mcmcOut[,"SchoolType-Public"], main = "95% HDI distribution of means of all vacci
nes in Public School")


# upper and lower bounds of 95% HDI
abline(v=quantile(mcmcOut[,"SchoolType-Public"], c(0.025)), col="blue")
abline(v=quantile(mcmcOut[,"SchoolType-Public"], c(0.975)), col="blue")
```

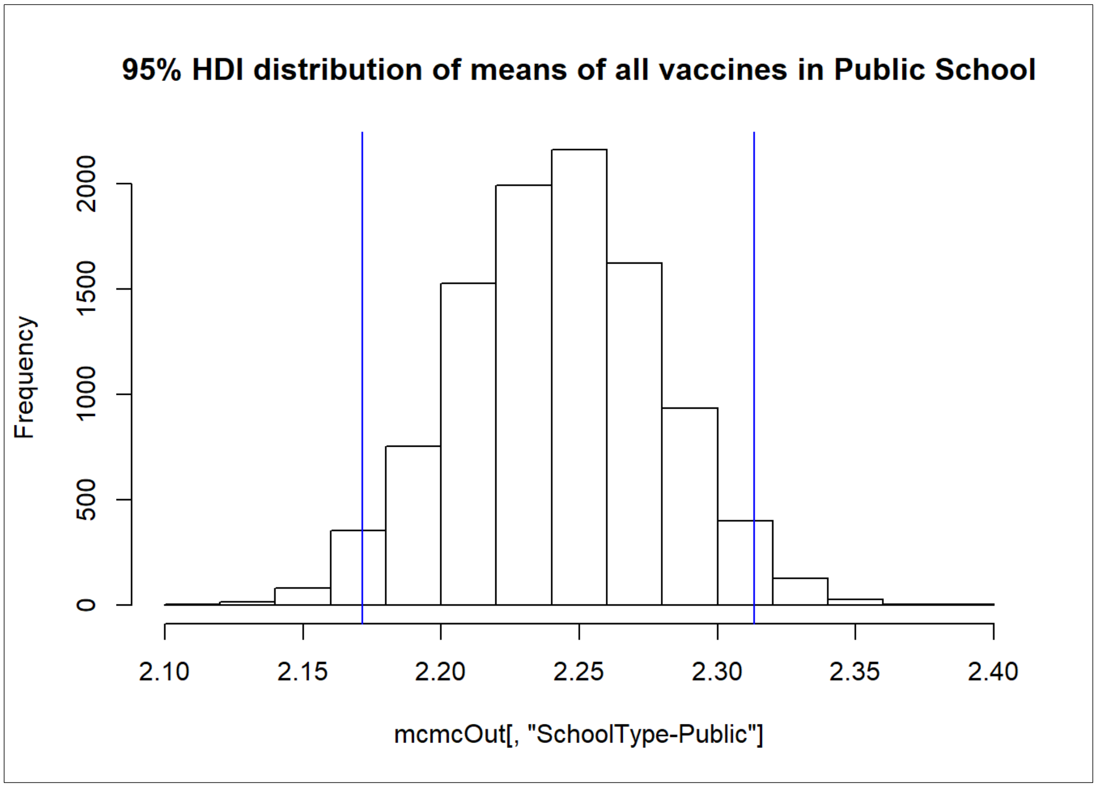


Figure 8: Distribution of Means of All Vaccines in Public Schools

The 95% HDI distribution does not include 0 as shown above which means it deviates from the population mean considerably. The 95% HDI upper bound to lower bounds ranges from 2.0860 to 2.231.

```
summary(mcmcOut)
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                        Mean        SD  Naive SE Time-series SE
## mu                   86.979 3.620e-02 3.620e-04      3.567e-04
## SchoolType-Public     2.242 3.631e-02 3.631e-04      3.631e-04
## SchoolType-Private -2.242 3.631e-02 3.631e-04      3.631e-04
## sig2                  2.613 8.163e-02 8.163e-04      8.163e-04
## g_SchoolType         41.296 1.677e+03 1.677e+01      1.677e+01
##
```

```
## 2. Quantiles for each variable:
##
##                         2.5%    25%    50%    75%  97.5%
## mu                  86.9066 86.955 86.979 87.004 87.049
## SchoolType-Public    2.1714  2.217  2.242  2.267  2.313
## SchoolType-Private -2.3134 -2.267 -2.242 -2.217 -2.171
## sig2                 2.4595  2.558  2.612  2.668  2.778
## g_SchoolType         0.5628  1.465  2.938  7.201 85.715
```

Bayesian Findings:

- From the summary data and the plots of 95% posterior probability distribution, the means of all vaccines for public and private schools are very different.
- In fact, the 95% HDI distributions do not overlap at all and each distribution does not straddle zero, showing they are further apart from the population mean.

Overall Conclusion:

- Overall, both frequentist ANOVA and Bayesian methods result in the same conclusion that the means of all vaccine rates between public and private schools are credibly different.

Compare medical exemptions between public and private schools. Are there any credible differences?

```
#str(reportSample)

#str(reportSample$medical)

# are there any missing values in allVaccs?

#length(complete.cases(reportSample$medical)) # no missing values

set.seed(123) # control the randomization

# obtain 100 random public school medical rates sample means, repeat X1000

medicalRates <- replicate(1000, mean(sample(PublicSchoolGroup$medical, 100, replace = TRUE)))

SchoolType <- as.factor("Public")

publicDF <- data.frame(medicalRates,SchoolType)

# obtain 100 random private school medical rates sample means, repeat X1000

medicalRates <- replicate(1000, mean(sample(PrivateSchoolGroup$medical, 100, replace = TRUE)))

SchoolType <- as.factor("Private")

privateDF <- data.frame(medicalRates,SchoolType)

# combine both groups into single group of both public and private medical rate samples

SchoolMeds <- rbind(publicDF,privateDF)

#View(SchoolMeds)
```

Figure 9 below shows the boxplot of the distributions of private and public sample means after random sampling process.

```
boxplot(medicalRates ~ SchoolType, data = SchoolMeds)
```

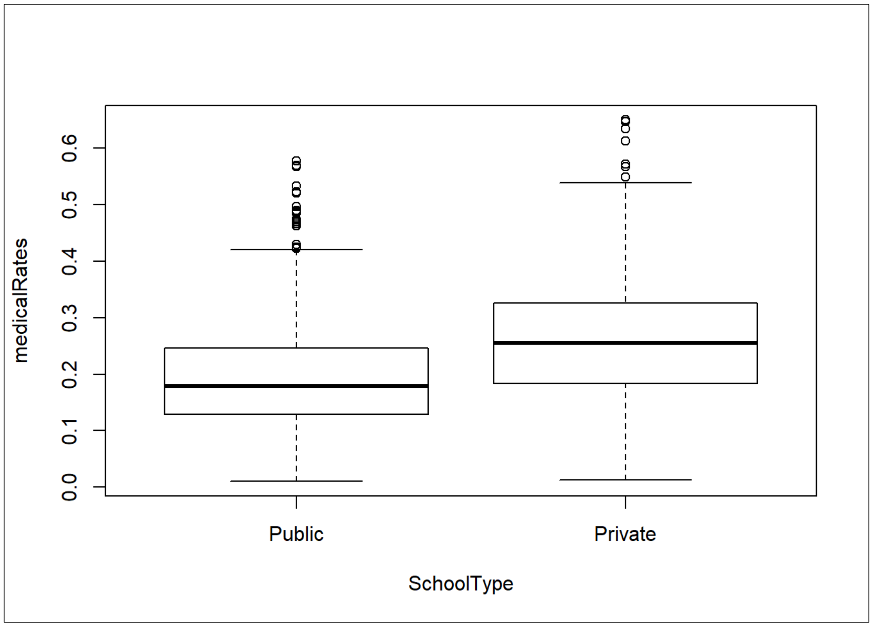


Figure 9: Boxplot of medical exemption sample means of Private and Public Schools

The boxplots show existing difference in median of Public vs Private school medical exemption rates, with medical exemption rates slightly higher in Private schools.

Frequentist Approach

```
vacOut <- aov(medicalRates ~ SchoolType, data = SchoolMeds)
summary(vacOut)
##               Df Sum Sq Mean Sq F value Pr(>F)
## SchoolType     1    1.819   1.8188 <2e-16 ***
## Residuals   1998 19.481  0.0098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There were a total of 2000 elements, 1000 each from public and private schools mean samples. The degrees of freedom residual elements are 1998 with 1 element left to vary between the two groups during the statistical calculation of ANOVA.
- The Sum Sq value is 1.819 between groups and 19.481 within groups - and it shows the raw initial calculation of variability.
- The Mean Sq is 1.8188 between groups and 0.0098 within groups - it is the variance (the sum of squares divided by the digress of freedom).
- The F-Value of the overall test is 186.5 and it is the ratio of the mean squares between and within groups.
- The Pr(>F) is <2e-16 which is the probability of a larger F-value of this ratio (186.5) within the F-distribution.

For the result to be statistically significant, the F-Value must substantially exceed 1 under the assumption that all the data was sampled from the same underlying population - as is the case here. Also, the Pr(>F) must be less than the alpha value (in this case, alpha is 0.05)

Frequentist Findings:

- The null hypothesis is that there is no difference in sample means of medical exemption rates between public and private schools.
- The alternative hypothesis is that the sample means of medical exemption rates between public and private schools are different.
- Since the Pr(>F) of <2e-16 is less than alpha-value (0.05), we reject the null hypothesis and conclude that the test is statistically significant; there is credible evidence that sample means of medicine rates between public and private schools are different.

Bayesian Approach

```
## Bayes factor analysis
## --------------
## [1] SchoolType : 5.731206e+43 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

The null hypothesis is that the difference in means between medical exemption rates in public and private schools is not different from zero (the population mean).

The alternative hypothesis is that the difference in means between medical exemption rates in public and private schools is significantly different from zero (the population mean).

The Bayes Factor analysis shows that there is a 5.731206e+43 to 1 odds in favor of the alternative hypothesis against an intercept only model.

The 95% HDI distribution of means of all medical exemption rates in Private School does not include 0 as shown below. It deviates from the population mean slightly. The 95% HDI upper bound to lower bounds ranges from 0.02570818 to 0.03436204.

```
par(mfcol=c(1,1))
hist(mcmcOut[,"SchoolType-Private"], main = "95% HDI dist: Medical Exemptions in Private School")


# upper and lower bounds of 95% HDI
abline(v=quantile(mcmcOut[,"SchoolType-Private"], c(0.025)), col="blue")
abline(v=quantile(mcmcOut[,"SchoolType-Private"], c(0.975)), col="blue")
```

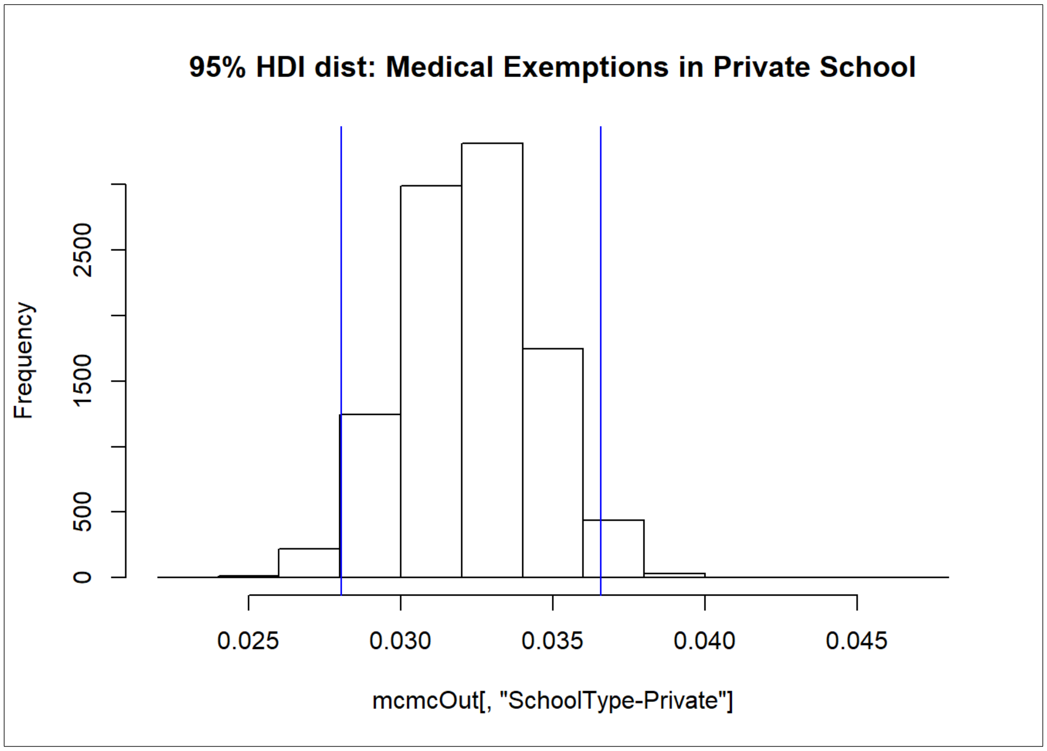


Figure 10: Medical Rates Sample Means in Private Schools

```
par(mfcol=c(1,1))

hist(mcmcOut[,"SchoolType-Public"], main = "95% HDI dist: Medical exemptions in Public
School")


# upper and lower bounds of 95% HDI

abline(v=quantile(mcmcOut[,"SchoolType-Public"], c(0.025)), col="blue")

abline(v=quantile(mcmcOut[,"SchoolType-Public"], c(0.975)), col="blue")
```

95% HDI distribution of means of all medical exemption rates in Public School does not include 0 as shown below. The 95% HDI upper bound to lower bounds ranges from -0.03436204 to -0.02570818.

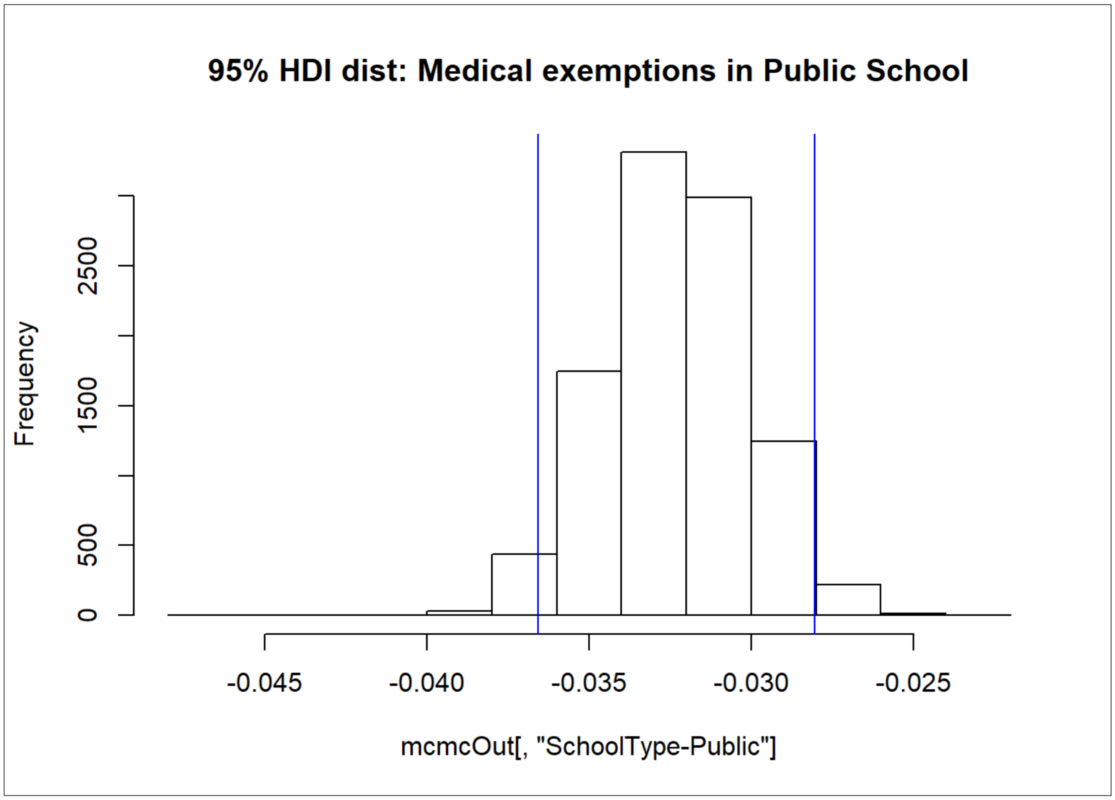


Figure 11: Medical Rates Sample Means in Public Schools

Summary outcomes

```
summary(mcmcOut)
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                          Mean        SD  Naive SE Time-series SE
## mu                   0.226780 2.226e-03 2.226e-05      2.189e-05
## SchoolType-Public   -0.032305 2.199e-03 2.199e-05      2.199e-05
## SchoolType-Private   0.032305 2.199e-03 2.199e-05      2.199e-05
## sig2                 0.009569 2.993e-04 2.993e-06      2.993e-06
## g_SchoolType         5.068875 2.256e+02 2.256e+00      2.256e+00
##
```

```
## 2. Quantiles for each variable:
##
##                          2.5%       25%       50%       75%    97.5%
## mu                  0.222408  0.225327  0.226796  0.228286  0.23100
## SchoolType-Public  -0.036576 -0.033810 -0.032282 -0.030808 -0.02805
## SchoolType-Private  0.028047  0.030808  0.032282  0.033810  0.03658
## sig2                0.009006  0.009366  0.009562  0.009771  0.01017
## g_SchoolType        0.064613  0.169126  0.335549  0.817836  9.78413
```

Overall Conclusion:

- From the summary data and the distribution plots, the means of medical exemptions for public and private schools are very different, in fact their 95% HDI distributions do not overlap at all and each does not straddle zero, showing they are apart from the population mean.
- Overall, both frequentist ANOVA and Bayesian methods result in the same conclusion that the means of medical exemption rates between public and private schools are credibly different.

Compare religious/belief exemptions between public and private schools. Are there any credible differences?

Data Prep and Sampling

```r
set.seed(123) # control the randomization

# obtain 100 random public school medical rates sample means, repeat X1000

religiousRates <- replicate(1000, mean(sample(PublicSchoolGroup$religious, 100, replac
e = TRUE)))

SchoolType <- as.factor("Public")

publicDF <- data.frame(religiousRates,SchoolType)

# obtain 100 random private school medical rates sample means, repeat X1000

religiousRates <- replicate(1000, mean(sample(PrivateSchoolGroup$religious, 100, repla
ce = TRUE)))

SchoolType <- as.factor("Private")

privateDF <- data.frame(religiousRates,SchoolType)

# combine both groups into single group of both public and private medical rate sample

SchoolRels <- rbind(publicDF,privateDF)

#View(SchoolMeds)
```

Figure 12 below shows the boxplot of the distributions of religious exemption rates sample means in private and public schools after random sampling process.

```r
boxplot(religiousRates ~ SchoolType, data = SchoolRels)
```

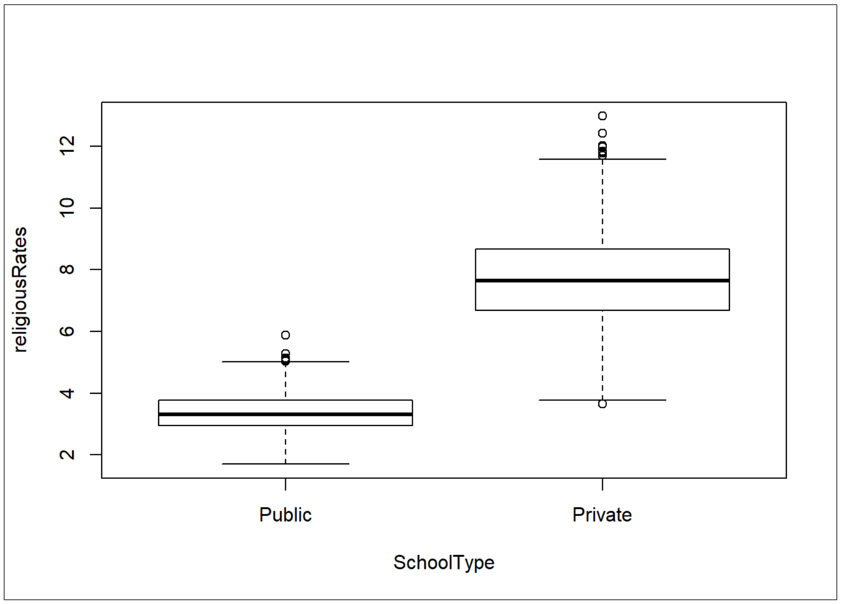


Figure 12 Religious Exemption Rates

The box plot shows that the medians of religious rates between public school and private schools are very different; private schools have a higher median.

The Frequentist Approach Using ANOVA

```
vacOut <- aov(religiousRates ~ SchoolType, data = SchoolRels)
summary(vacOut)
##               Df Sum Sq Mean Sq F value Pr(>F)
## SchoolType     1   8862    8862    7203 <2e-16 ***
## Residuals   1998   2458       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There were a total of 2000 elements, 1000 each from public and private schools. The degrees of freedom residual elements are 1998 with 1 element left to vary between the two groups during the statistical calculation of ANOVA.
- The Sum Sq value is 8862 between groups and 2458 within groups - and it shows the raw initial calculation of variability.
- The Mean Sq is 8862 between groups and 1 within groups - and it is the variance (the sum of squares divided by the degress of freedom).
- The F-Value of the overall test is 7203 and it is the ratio of the mean squares between and within groups.
- The Pr(>F) is <2e-16 - it is the probability of a larger F-value of this ratio (7203) within the F-distribution.

For the result to be statistically significant, the F-Value must substantially exceed 1 under the assumption that all the data was sampled from the same underlying population - as is the case here. Also, the Pr(>F) has to be less than the alpha value (in this case, alpha threshold is $p < 0.05$)

Frequentist Findings:

- The null hypothesis is that there is no difference in sample means of religious rates between public and private schools.
- The alternative hypothesis is that the sample means of religious rates between public and private schools are different.

- Since the Pr(>F) of <2e-16 is less than alpha-value (0.05), we reject the null hypothesis and conclude that the test is statistically significant; there is credible evidence that sample means of religious rates between public and private schools are different.

Bayesian Approach using Bayes Factor Analysis

```
vacBayesOut <- anovaBF(religiousRates ~ SchoolType, data = SchoolRels)
summary(vacBayesOut)
## Bayes factor analysis
## --------------
## [1] SchoolType : 1.816783e+683 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

The Bayes Factor of 1.816783e+683 is strong evidence to the alternative hypothesis, that the sample means of religious rates are different between public and private schools.

The 95% HDI distribution of means of religious exemption rates in Private School**s** does not include 0 as shown below which means it deviates from the population mean considerably. The 95% HDI upper bound to lower bounds ranges from 2.0557 to 2.153556. This is illustrated below in *Figure 13*.

```
par(mfcol=c(1,1))
hist(mcmcOut[,"SchoolType-Private"], main = "95% HDI distribution of means of Religious rates in Private School")

# upper and lower bounds of 95% HDI
abline(v=quantile(mcmcOut[,"SchoolType-Private"], c(0.025)), col="blue")
abline(v=quantile(mcmcOut[,"SchoolType-Private"], c(0.975)), col="blue")
```

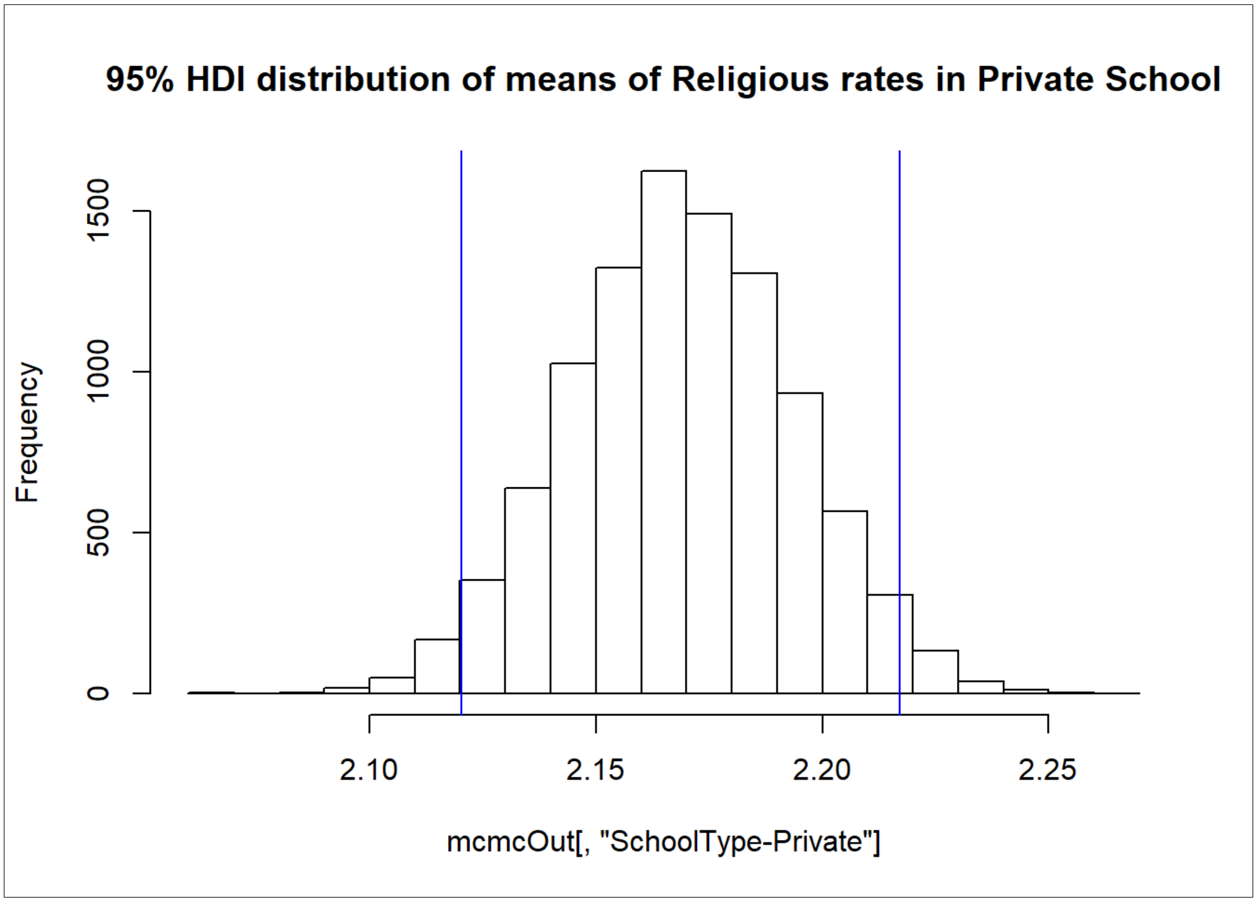


Figure 13: Distribution of Mean Religious exemption rates in Private Schools

The 95% HDI distribution of means of religious exemption rates in Public School **does not include 0** as shown below which means it deviates from the population mean considerably. The 95% HDI upper bound to lower bounds ranges from -2.153556 to -2.0557. *Figure 14* below gives a visual rendition of this finding.

```
par(mfcol=c(1,1))

hist(mcmcOut[,"SchoolType-Public"], main = "95% HDI distribution of means of
Religious Rates in Public School")


# upper and lower bounds of 95% HDI

abline(v=quantile(mcmcOut[,"SchoolType-Public"], c(0.025)), col="blue")

abline(v=quantile(mcmcOut[,"SchoolType-Public"], c(0.975)), col="blue")
```

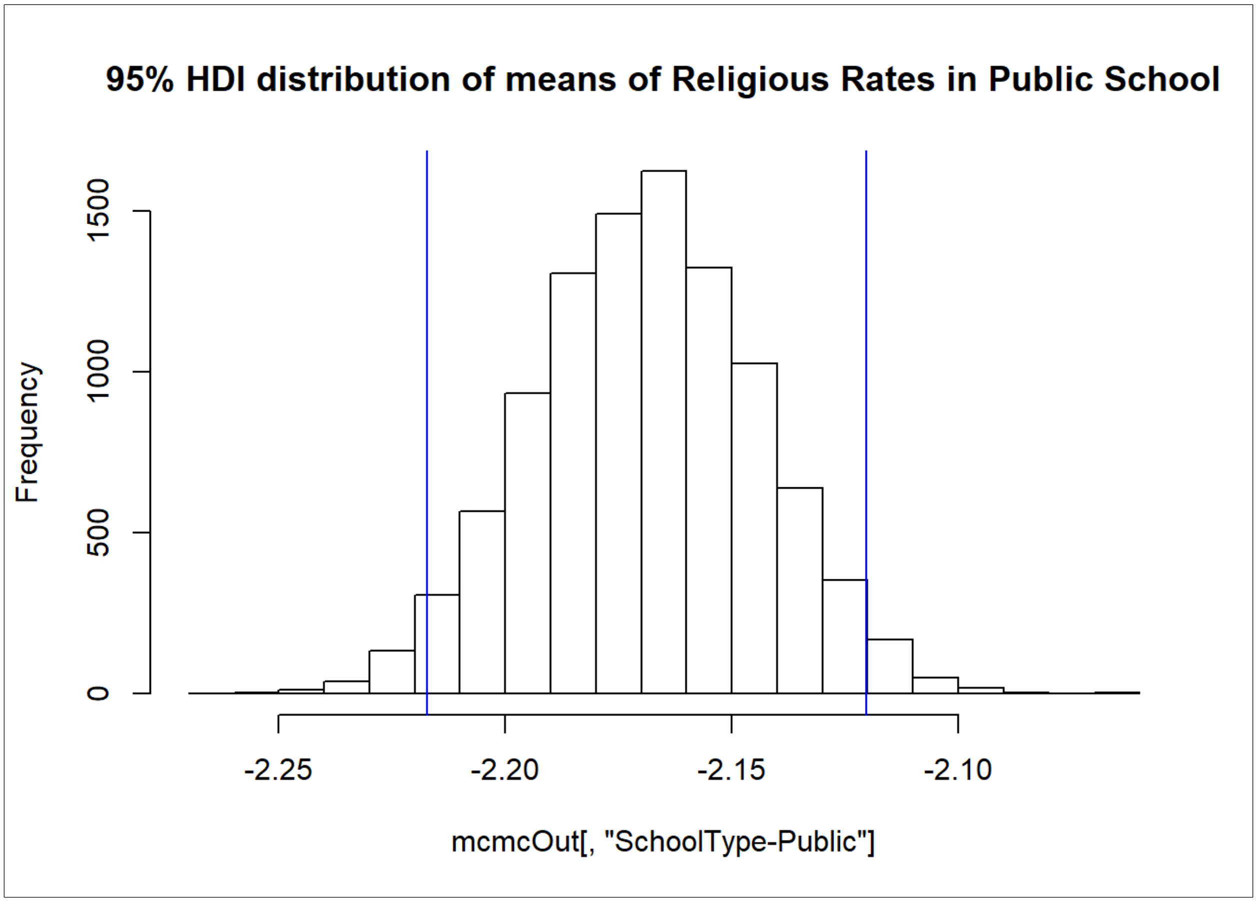
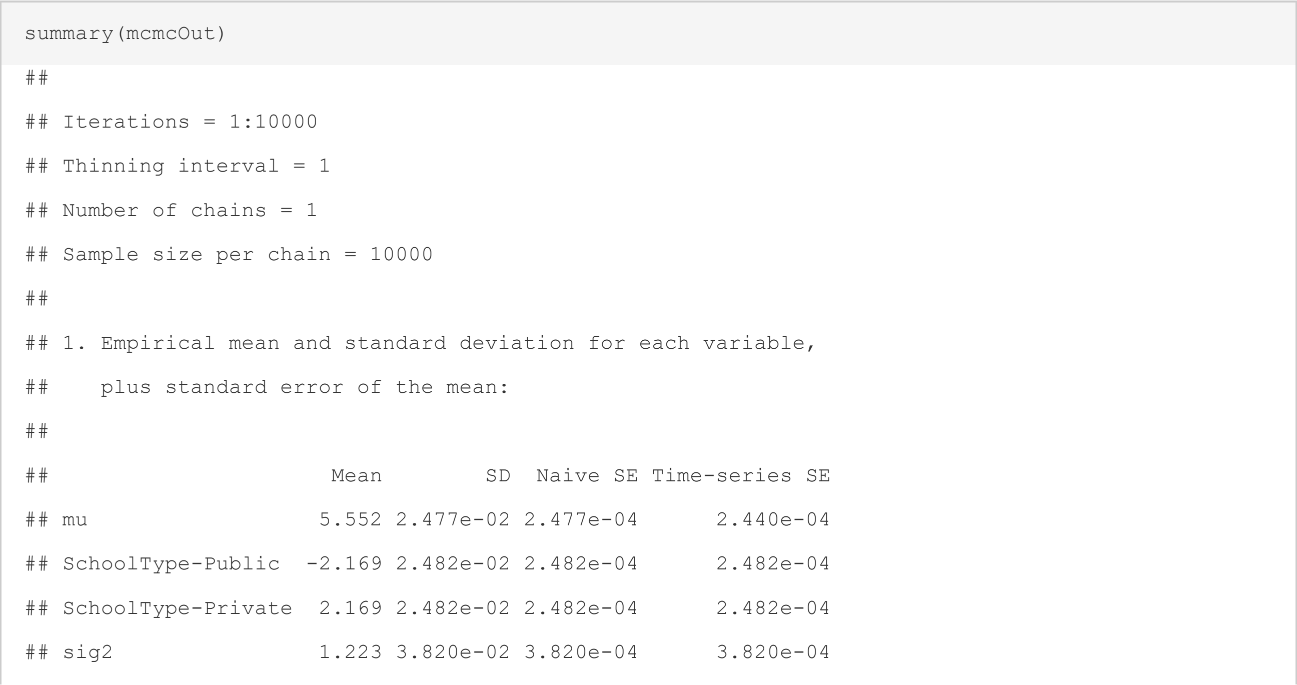


Figure 14: Distribution of Mean Religious exemption rates in Public Schools



```
summary(mcmcOut)
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                        Mean        SD  Naive SE Time-series SE
## mu                    5.552 2.477e-02 2.477e-04      2.440e-04
## SchoolType-Public    -2.169 2.482e-02 2.482e-04      2.482e-04
## SchoolType-Private    2.169 2.482e-02 2.482e-04      2.482e-04
## sig2                  1.223 3.820e-02 3.820e-04      3.820e-04
## g_SchoolType       82.928 3.542e+03 3.542e+01      3.542e+01
##
## 2. Quantiles for each variable:
##
##                       2.5%    25%    50%    75%   97.5%
## mu                   5.503  5.536  5.552  5.569   5.600
## SchoolType-Public  -2.217 -2.186 -2.169 -2.152  -2.120
## SchoolType-Private  2.120  2.152  2.169  2.186   2.217
## sig2                 1.151  1.197  1.222  1.249   1.300
## g_SchoolType         1.097  2.859  5.687 13.921 166.040
```

Bayesian Finding:

From the summary data and the distribution plots, the means of religious rates for public and private schools are very different, in fact their 95% HDI distributions do not overlap at all and each is does not straddle zero, showing they are further apart from the population mean.

Overall Conclusion:

Overall, both frequentist ANOVA and Bayesian methods result in the same conclusion that the means of all vaccine rates between public and private schools are credibly different.

PREDICTIVE ANALYSES

Is it possible to predict whether a school is public or private based on conditional, medical, and religious percentages? If so, what are the specifics?

Frequentist Approach

Building the Model and Interpreting the Model Summary

```
chOut <- glm(formula = pubpriv ~ conditional + medical + religious, family = binomial(
), data = reportSample)


summary(chOut)
##
## Call:
## glm(formula = pubpriv ~ conditional + medical + religious, family = binomial(),
##     data = reportSample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9492   0.5693   0.5841   0.6181   1.7478
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.737610   0.133753  12.991  < 2e-16 ***
## conditional -0.002034   0.008838  -0.230    0.818
## medical     -0.047837   0.093419  -0.512    0.609
## religious   -0.043794   0.010287  -4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.17  on 697  degrees of freedom
## Residual deviance: 644.69  on 694  degrees of freedom
## AIC: 652.69
##
## Number of Fisher Scoring iterations: 4
```

A general linear model is used that takes conditional, medical and religious percentages as the predictor variables to predict the school type (Private or Public) The general linear model uses the binomial family for logistic output.

Frequentist Findings:

- The output shows that the intercept is significantly different from zero - since the intercept represents the log-odds of a "Public" outcome when all the predictor variables are equal to zero.
- The coefficient of the conditional rate predictor is not statistically significant based on the Wald's z-value of -0.23 and associated p-value of 0.818. Since p = 0.818 means p > .001, we fail to reject the null hypothesis that the log-odds of conditional rates is 0 in the population.
- The coefficient of the medical rate predictor is not statistically significant based on the Wald's z-value of -0.51 and associated p-value of 0.609. Since p = 0.609 means p > .001, we fail to reject the null hypothesis that the log-odds of medical rates is 0 in the population.
- The coefficient of the religious rate predictor is statistically significant based on the Wald's z-value of -4.26 and associated p-value of 2.07e-05. Since p = 2.07e-05 means p < .001, we reject the null hypothesis that the log-odds of conditional rates is 0 in the population.

Converting log odds to straight odds

```
coef(chOut)
##  (Intercept)  conditional      medical    religious
##  1.737609870 -0.002034471 -0.047837183 -0.043793641
```

The straight odds show that the intercept represent odds of 1.74:1 for the school type being "Private" given that all the predictors are equal to zero. The odds of -0.002:1 for conditional rate show that for every conditional exemption, the school is 0.2% less likely to be a private school. In the case of medical rates, the odds are -0.05:1, meaning that the odds of being a private school decreases with every increase in medical exemption rate by approximately 0.5%. In the case of religious rates, the odds are -0.05:1, meaning that the odds of being a private school decreases with every increase in medical exemption rate by approximately 0.5%.

Examination of confidence intervals

```
exp(confint(chOut))
## Waiting for profiling to be done...
##                 2.5 %    97.5 %
## (Intercept) 4.3960352 7.4304891
## conditional 0.9816586 1.0165535
## medical     0.8008185 1.1745340
## religious   0.9369964 0.9758728
```

The 95% confidence interval for religious predictor rates ranges from a low of 0.9369964:1 to a high of 0.9758728:1. At the low end of the confidence interval, an increase in religious exemption rates by 1% corresponds to 0.9369964:1 in favor a Private school while at the high end of the confidence interval, 1% increase in religious exemptions corresponds to a 0.9758728:1 in favor this being a private school.

Reporting the Chi-square test:

```
anova(chOut, test = "Chisq")
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: pubpriv
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          697     665.17
## conditional  1   0.0043       696     665.16    0.9478
## medical      1   0.4445       695     664.72    0.5049
## religious    1  20.0269       694     644.69 7.636e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test contains three nested models, with the first chi-square test comparing the null model to a model with just the conditional rates, the second one contains medical and religious, and the third one contains just religious.

Only the last chi-square test is statistically significant because p = 7.636e-06 is below the threshold of p < 0.001. This outcome makes sense in light of the significance tests on the coefficients and confirms the utility of a model that contains only religious predictor variables.

Tabulating “Private” vs “Public” Predictions

```
table(round(predict(chOut, type = "response")), reportSample$pubpriv)
##
##      PRIVATE PUBLIC
##    0       7      4
##    1     121    566
```

There were 7 cases where the observed school was Private, and the dichotomized predictor value was “Private”. These are the correct classification of “Private” schools. Likewise, there were 566 cases where the observed school was Public, and the dichotomized predictor value was “Public”.

Thus, the overall accuracy of the model is (7 + 566)/698 which is 0.82 or 82% accuracy. 698 denominator is the total of all observations.

Examining the Pseudo-R Value

```
library(BaylorEdPsych)
PseudoR2(chOut)
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##       0.03078278       0.01574900       0.02890878       0.04705179
## McKelvey.Zavoina           Effron            Count        Adj.Count
##       0.04246562       0.03766105       0.82091691       0.02343750
##              AIC    Corrected.AIC
##     652.69292814     652.75064820
```

The Nagelkerke pseudo-R is reports the highest value indicating the proportion of variance in the outcome variable (privpub) that accounted for by the predictor variables (religious, medical, conditional). However, this value is very small which means that since only religious rates were significant in the prediction model, it accounts for a minimal proportion of that variation in the outcome. This contrast maybe attributed to statistical power of the sample size assessed (698 observations) which created this relatively small effect vis-a-viz size.

Bayesian Approach

Data Prep

The Bayes output requires numeric input; thus, a transformation on the school type (pubpriv) is done as shown below:

```
# adjust the outcome variable
reportSample$pubpriv <- as.numeric(reportSample$pubpriv) - 1
```

Creating the Bayes Logit Output

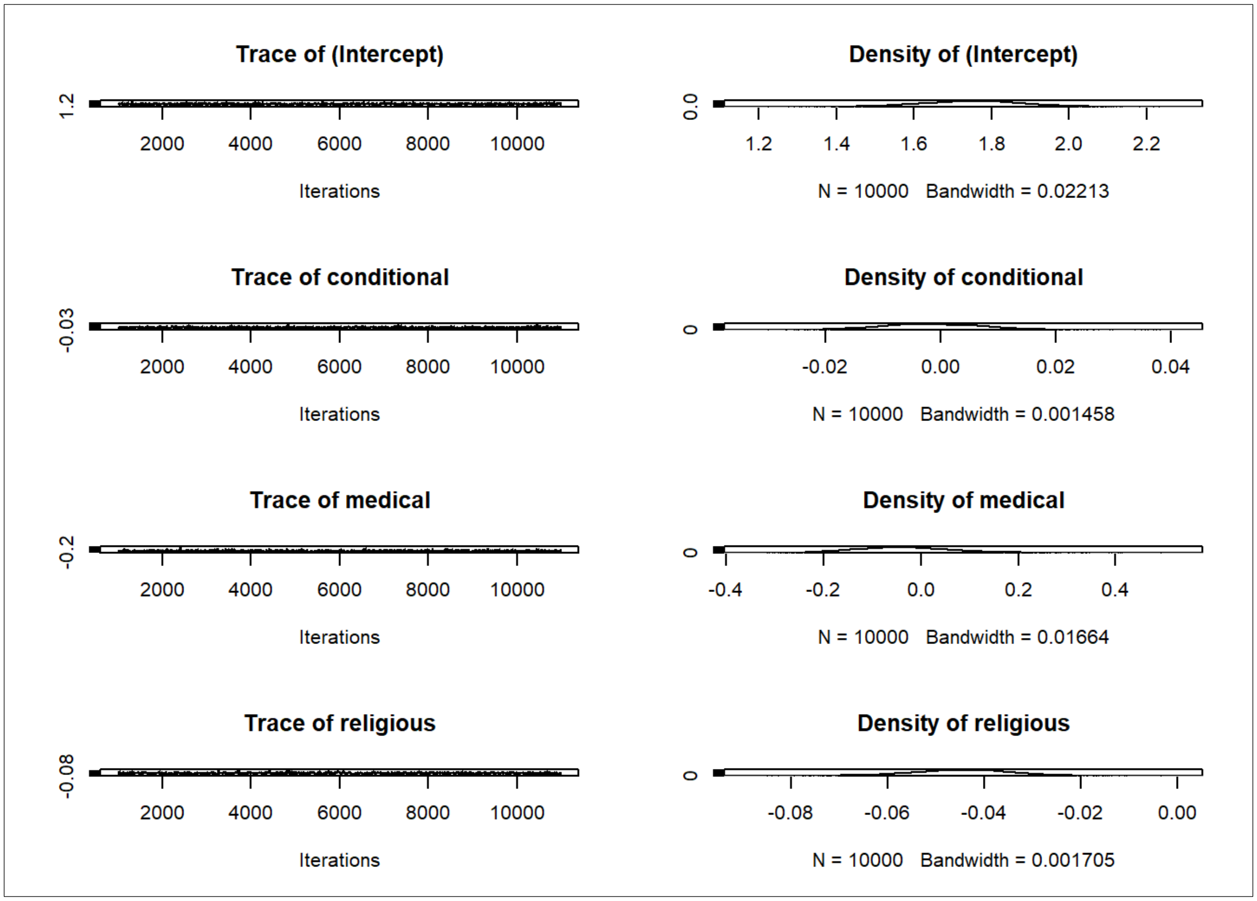
```
bayesLogitOut <- MCMClogit(formula = pubpriv ~ conditional + medical + religious, data
= reportSample)


summary(bayesLogitOut)
```

```
##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                  Mean       SD  Naive SE Time-series SE
## (Intercept)  1.749391 0.133244 1.332e-03      0.0048761
## conditional -0.001228 0.008678 8.678e-05      0.0003227
## medical     -0.032916 0.100538 1.005e-03      0.0038818
## religious   -0.044844 0.010425 1.042e-04      0.0003785
##
## 2. Quantiles for each variable:
##
##                 2.5%      25%       50%       75%    97.5%
## (Intercept)  1.48348  1.66130  1.752757  1.837843  2.00639
## conditional -0.01736 -0.00722 -0.001633  0.004734  0.01596
## medical     -0.20789 -0.10321 -0.040479  0.029548  0.18424
## religious   -0.06658 -0.05154 -0.044682 -0.037940 -0.02539
```

The summary outcome focuses on describing the posterior distributions of parameters representing both the intercept and the coefficients on conditional, medical and religious, calibrated as log-odds.

The point estimate of the intercept and the coefficients are very close or similar to the frequentist approach described earlier.

A plot of the region between the 2.5% and 97.5% quantiles is the are of High Density Interval as shown in the plots below:



The left Colum of the plot gives the trace of the progress of the *MCMClogit()* function as it performs the Markov chain Monte Carlo analysis.

The right column shows the graphical representations of the likely position of each coefficient. The true population value of each coefficient is likely to be somewhere in the middle of the distribution and less likely near the tails.

The density plots of conditional, and medical overlaps with zero, while the distribution of religious does not. This is a little unexpected considering that the frequentist approach showed religious variable to be the only significant variable in predicting the model.

Examining Regular Odds for Conditional Percentages

```
# creat a matrix for apply()
conditionalLogOdds <- as.matrix(bayesLogitOut[,"conditional"])
# apply() runs exp() for each one
conditionalOdds <- apply(conditionalLogOdds, 1, exp)
# plot hist
hist(conditionalOdds)
# left edge of 95% HDI
abline(v=quantile(conditionalOdds, c(0.025)), col="blue")
# right edge of 95% HDI
abline(v=quantile(conditionalOdds, c(0.975)), col="blue")
```

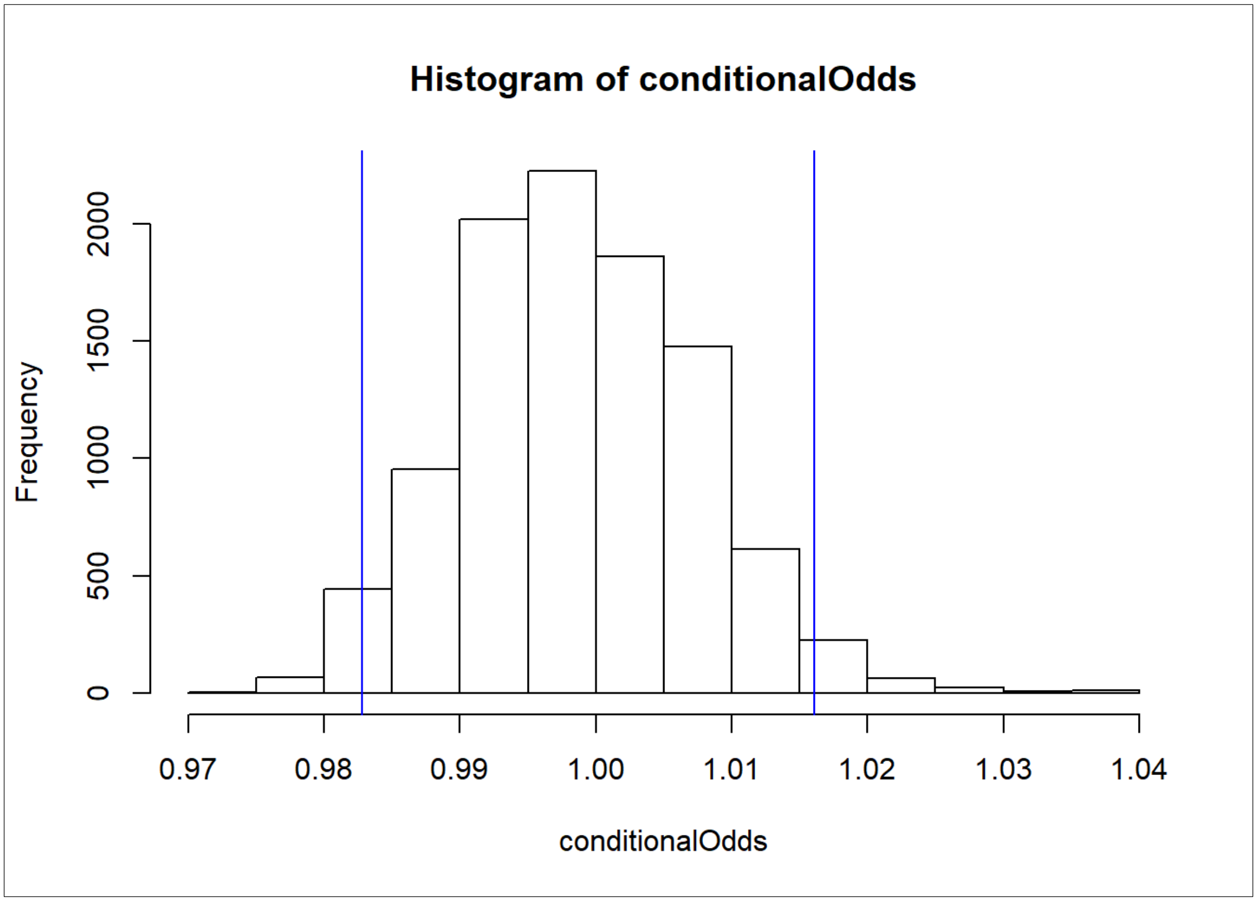


Figure 15: Mean is centered around 1

Examining the Log Odds for Religious Percentages

```
# creat a matrix for apply()
religiousLogOdds <- as.matrix(bayesLogitOut[,"religious"])
# apply() runs exp() for each one
religiousOdds <- apply(religiousLogOdds, 1, exp)
# plot hist
hist(religiousOdds)
# left edge of 95% HDI
abline(v=quantile(religiousOdds, c(0.025)), col="blue")
# right edge of 95% HDI
abline(v=quantile(religiousOdds, c(0.975)), col="blue")
```

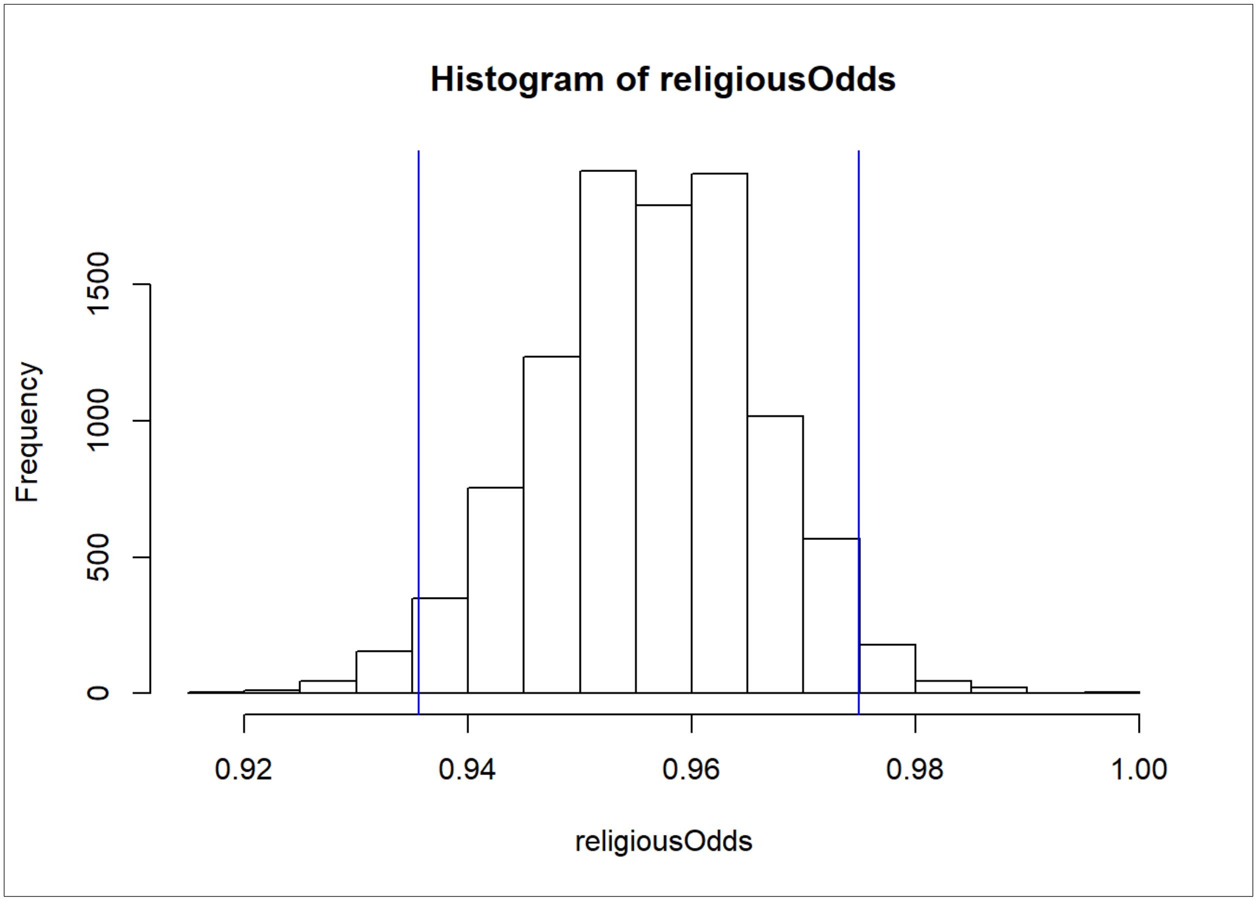
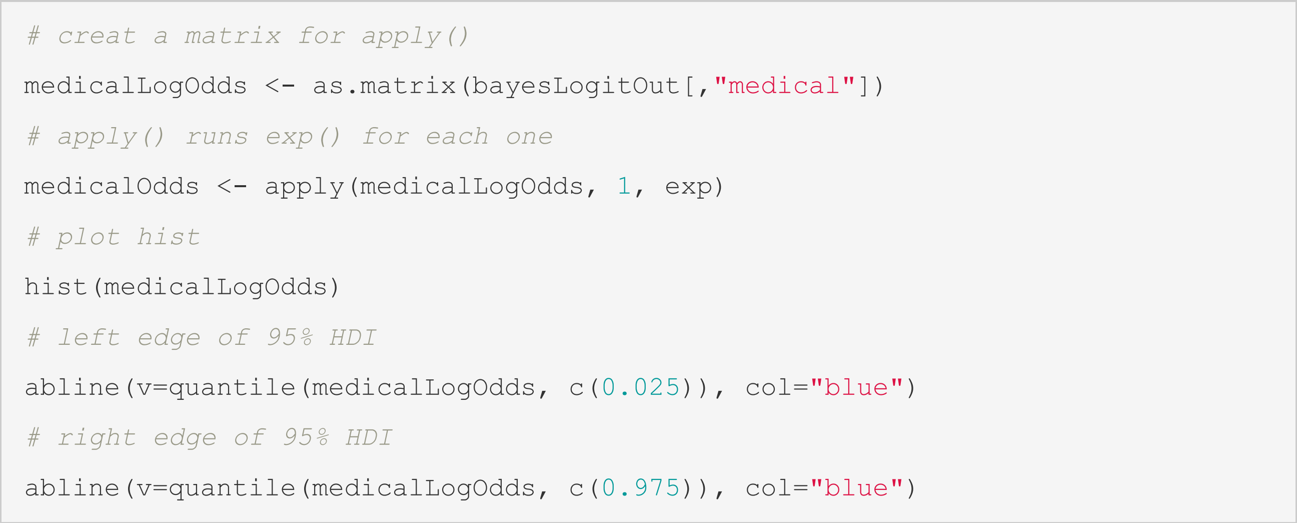


Figure 16: Centered around 0.96.

Examining the Log Odds for Medial Percentages



```
# creat a matrix for apply()
medicalLogOdds <- as.matrix(bayesLogitOut[,"medical"])
# apply() runs exp() for each one
medicalOdds <- apply(medicalLogOdds, 1, exp)
# plot hist
hist(medicalLogOdds)
# left edge of 95% HDI
abline(v=quantile(medicalLogOdds, c(0.025)), col="blue")
# right edge of 95% HDI
abline(v=quantile(medicalLogOdds, c(0.975)), col="blue")
```

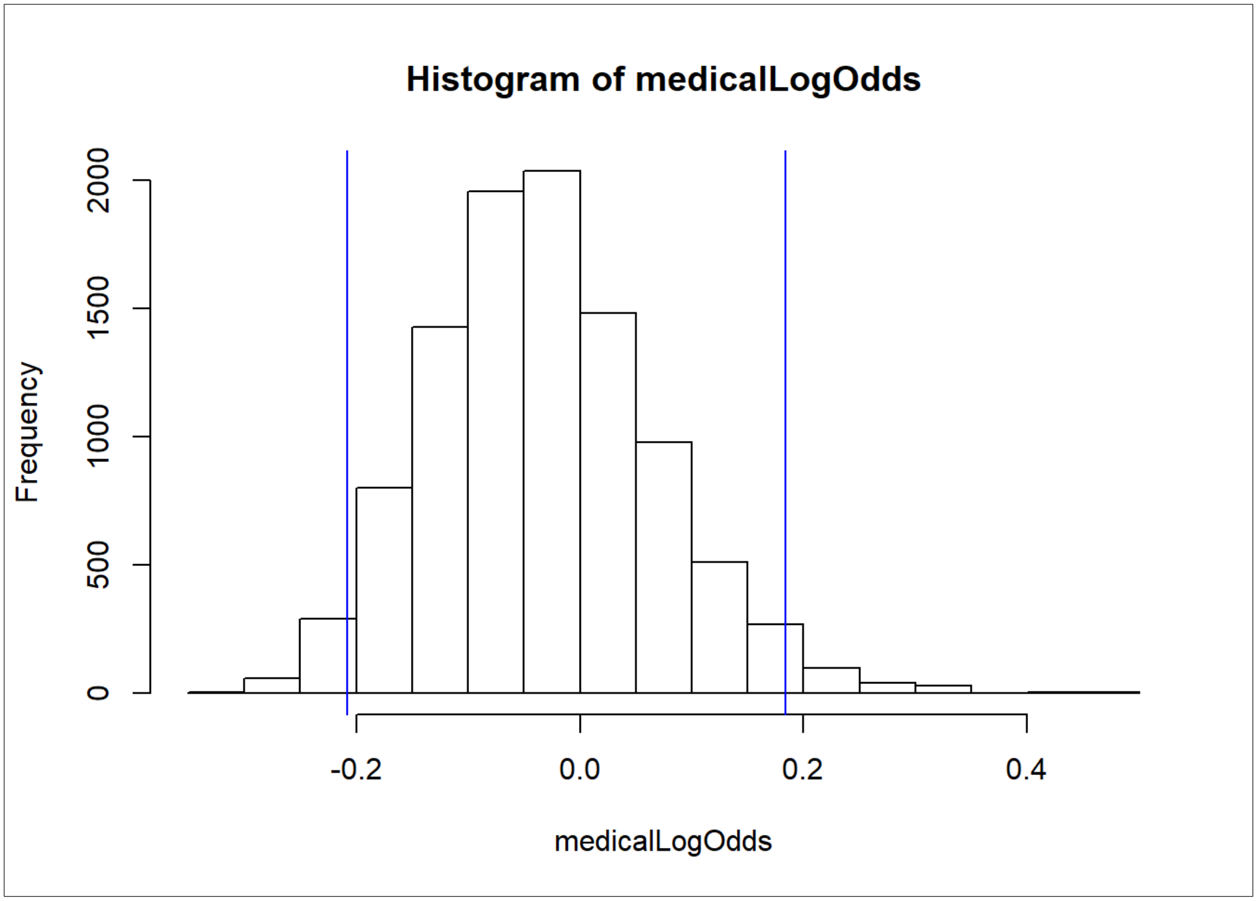


Figure 17: Mean is Centered around 0

Conclusion:

The 95% HDI distribution of straight odds show that medical and conditional rates are at nearly 1:1 odds while religious rates are at 0:1 odds. This is strange and differs significantly from the findings in the frequentist analysis.

Due to the different outcomes from the frequentist and Bayesian approaches, it is inconclusive whether a prediction of school type is possible based on conditional, religious or medical exemption rates.

Is it possible to predict conditional percentage, based on the percentages of specific vaccines that are missing? If so, what are the specifics?

Frequentist Approach

Build the Prediction Model and Evaluate Model Summary

The dependent variable is the conditional percentages, with the missing vaccines as the independent variables.

```
predVac <- lm(conditional ~ dptMiss + polMiss + mmrMiss + hepMiss + varMiss, data = reportSample)
summary(predVac)
```

```
## 
## Call:
## lm(formula = conditional ~ dptMiss + polMiss + mmrMiss + hepMiss + 
##     varMiss, data = reportSample)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -25.080  -2.070  -0.669   0.623  68.458 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.66856    0.32232   2.074  0.03843 *  
## dptMiss      0.71746    0.09338   7.683 5.33e-14 ***
## polMiss      0.34005    0.10505   3.237  0.00127 ** 
## mmrMiss      0.39540    0.06502   6.081 1.97e-09 ***
## hepMiss     -0.03585    0.09923  -0.361  0.71799    
## varMiss     -1.10833    0.10025 -11.056  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.603 on 692 degrees of freedom
## Multiple R-squared:  0.6502, Adjusted R-squared:  0.6476
## F-statistic: 257.2 on 5 and 692 DF,  p-value: < 2.2e-16
```

- The summary output shows the formula applied in the multivariate linear regression model.
- It then shows the residual summaries, and the intercept together with the B-weights (co-efficient) summaries that describe the line of best fit that reduces the sum of squared errors the most.
- The intercept estimate is 0.66856 denoting the value of the predicted outcome (conditional percentages) when all the missing vaccine values are equal to zero.
- The first three coefficients are very similar and are positive (0.71, 0.34, 0.39) except for the last two that are negative (-0.04, -1.11) The coefficient denote the contribution to the outcome, hence the first three missing vaccines (dptMiss, polMiss, mmrMiss) have a positive contribution to the dependent variable while hepMiss and varMiss have negative contribution to the dependent variable.
- The t-values and the probabilities associated with those t-values test the null hypothesis that each of the coefficients are equal to zero. In this case, we can reject the null hypothesis because of the small t-values and even smaller p-values for all the independent variables except in hepMiss.

Re-run the model after pruning

Since hepMiss has a p-value that is not significant, p-value = 0.71799, which is higher than alpha at p < 0.05, we drop it from the equation and rerun the model:

```
predVac <- lm(conditional ~ dptMiss + polMiss + mmrMiss + varMiss, data = reportSample)
summary(predVac)
```

```
##
## Call:
## lm(formula = conditional ~ dptMiss + polMiss + mmrMiss + varMiss,
##     data = reportSample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.123  -2.077  -0.655   0.602  68.400
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.65548    0.32008   2.048  0.04095 *
## dptMiss      0.71917    0.09321   7.716 4.19e-14 ***
## polMiss      0.33356    0.10344   3.225  0.00132 **
## mmrMiss      0.39583    0.06497   6.093 1.84e-09 ***
## varMiss     -1.13912    0.05279 -21.578  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 6.599 on 693 degrees of freedom
## Multiple R-squared:  0.6501, Adjusted R-squared:  0.6481
## F-statistic: 321.9 on 4 and 693 DF,  p-value: < 2.2e-16
```

Now, all p-values of the coefficients are less than alpha threshold $p < 0.05$.

The model summary shows the multiple R-squared value of 0.6501 which shows well over half of the variation in conditional percentages can be accounted for by the four predictor missing vaccine variables working together.

The Null hypothesis Significance test (F-test) is whether the R-squared value is significantly different from zero. In this case we reject the null hypothesis, because the model p-value of < 2e-16 is much less than the alpha threshold of $p < 0.05$. The test is statistically significant.

Thus, the prediction of conditional percentages based on missing vaccines is as follows, where the brackets denote multiplication:

Conditional Rate = 0.65548 + 0.71917(dptMiss) + 0.33356(polMiss) + 0.39583(mmrMiss) - 1.13912(varMiss)

Analyzing Residuals for normality/non-normality

The mean of the residuals from the least squares processing should always be zero.

```
hist(predVac$residuals)
```

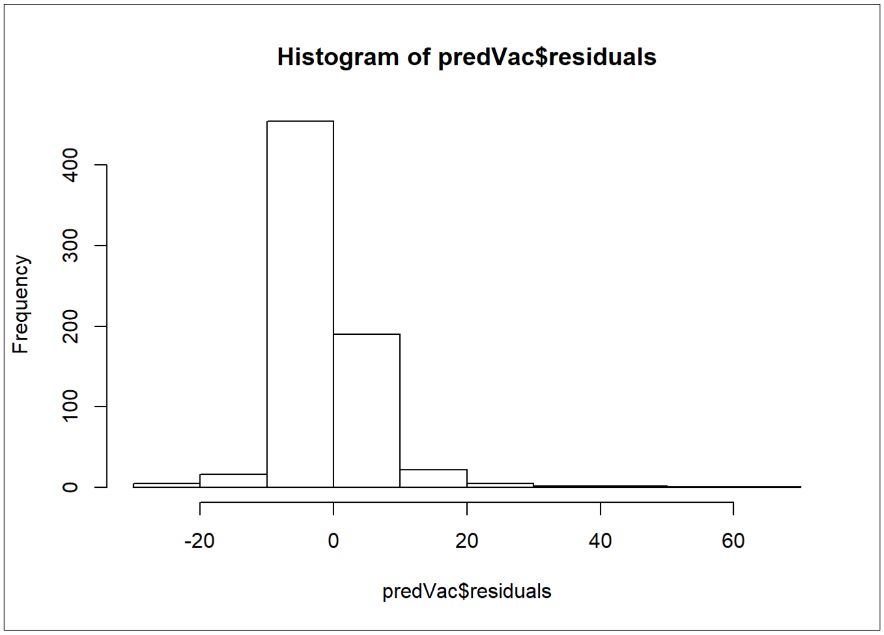


Figure 18 Residual Plot

The plot shows a largely normal distribution with the mean centered around zero indicating that the underlying relationship between the independent variables and the dependent variables are linear.

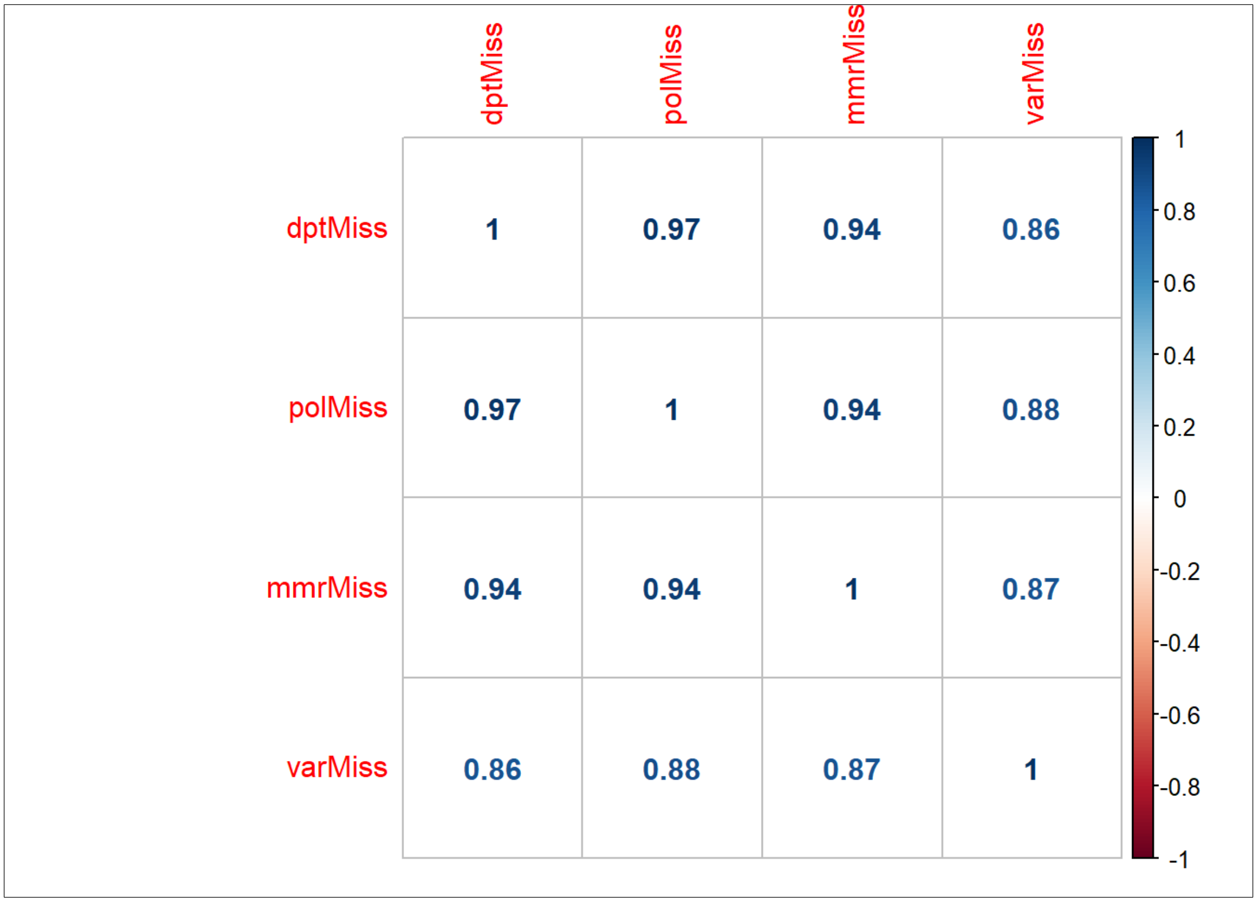
Testing for multicollinearity

By examining the Variance Inflation Factors for the independent variables as a diagnostic for multicollinearity. Whenever two independent variables are correlated, the slope (B-Weight) that is calculated only reflects the independent influence on the dependent variable, with the correlated part invisible on the weights.

```
library(car)
## Warning: package 'car' was built under R version 3.6.3
## Loading required package: carData
vif(predVac)
##   dptMiss   polMiss   mmrMiss   varMiss
## 20.561278 24.815952 10.288732  4.728519
```

High VIF values indicate multicollinearity. All the independent variables show a VIF value greater than 3 which are indicative of possible multicollinearity. This is best confirmed by the correlation plot below with corresponding correlation strengths of the independent variables.

```
# dataframe of missing vaccines
missingVacs <- data.frame(reportSample$dptMiss, reportSample$polMiss, reportSample$mmrMiss, repor
tSample$varMiss)


# label the missing vacciness
titles <- c("dptMiss", "polMiss", "mmrMiss", "varMiss")
colnames(missingVacs) <- titles


# run the correlation
my_cor <- cor(missingVacs)
corrplot(my_cor, method = "number")
```



	dptMiss	polMiss	mmrMiss	varMiss
dptMiss	1	0.97	0.94	0.86
polMiss	0.97	1	0.94	0.88
mmrMiss	0.94	0.94	1	0.87
varMiss	0.86	0.88	0.87	1

Conclusion:

As shown above, the independent variables exhibit very strong correlations with each other. This can take away from the strength of the model. The viability of the R-Squared value of 0.6501 can be improved by removing multicollinearity.

Bayesian Approach

Build the Prediction Model and Evaluate Model Summary

```
regOutMCMC <- lmBF(conditional ~ dptMiss + polMiss + mmrMiss + varMiss, data = reportSample, posterior=TRUE, iterations = 10000)


summary(regOutMCMC)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##             Mean      SD  Naive SE Time-series SE
## mu        7.2479 0.25217 0.0025217      0.0024785
## dptMiss   0.7178 0.09246 0.0009246      0.0009246
## polMiss   0.3302 0.10298 0.0010298      0.0010298
## mmrMiss   0.3945 0.06500 0.0006500      0.0006500
## varMiss  -1.1344 0.05327 0.0005327      0.0005341
## sig2     43.7449 2.35544 0.0235544      0.0235544
## g         0.6596 0.79319 0.0079319      0.0081313
##
## 2. Quantiles for each variable:
##
##             2.5%     25%     50%     75%   97.5%
## mu        6.7557  7.0804  7.2479  7.4163  7.7373
## dptMiss   0.5380  0.6552  0.7171  0.7796  0.9000
## polMiss   0.1247  0.2616  0.3305  0.3993  0.5292
## mmrMiss   0.2676  0.3510  0.3947  0.4379  0.5222
## varMiss  -1.2385 -1.1702 -1.1339 -1.0990 -1.0301
## sig2     39.3734 42.1389 43.6607 45.2704 48.6080
## g         0.1546  0.2991  0.4484  0.7280  2.4941
```

- The output shows the means of the respective distributions and the 95% HDIs. The mean colum are the paremeter estimates of the B-Weights of each of the predictions and they are a very close match to the frequentist calculations prior.
- The 2.5% and 97.5% boundaries of 95% HDIs for the IVs are very similar exept for hepMiss - this reflects similar finding in the frequentist approach where hepMiss was dropped from the initial model.
- The sig2(sigma-squared) shows the model precision for each of the iterations. The smaller sig2 is, the better the prediction.

Converting Sigma-squared values to R-squared values

The R-squared value of each model in the posterior distribution is equal to 1 minus the value of sig2 divided by the variance in the dependent variable.

```
rsqList <- 1 - (regOutMCMC[,"sig2"] / var(reportSample$conditional))
# mean
mean(rsqList)
## [1] 0.6464755
hist(rsqList)
# lower bound of 95% HDI
abline(v=quantile(rsqList, c(0.25)), col="blue")
# upper bound of 95% HDI
abline(v=quantile(rsqList, c(0.975)), col="blue")
```

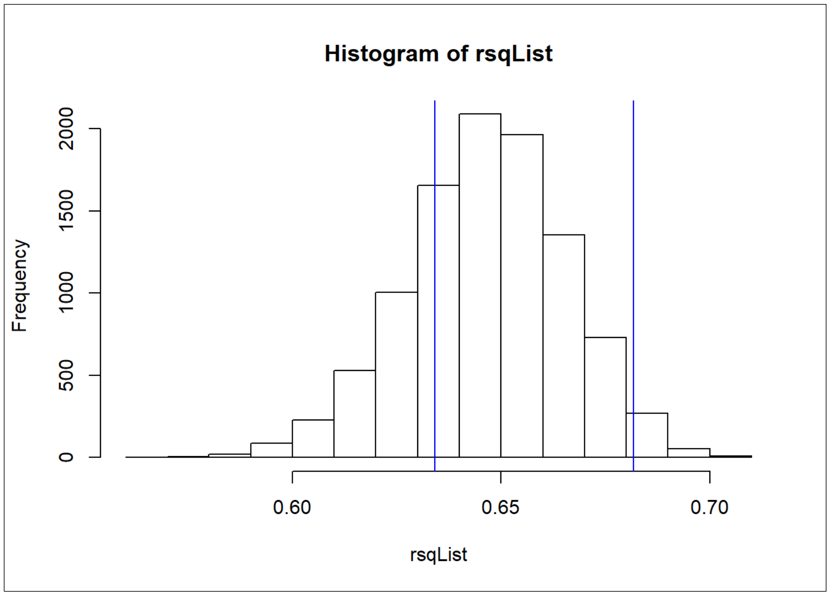


Figure 19: Pseudo-R Distribution

Conclusions

The mean value of the distribution came to 0.65 which is exactly similar to the R-squared value obtained using the frequentist approach prior. The Bayesian model also presents a clear-eyed view of the likely range of possibilities for the predictive strength of the model. It is possible to expect an R-squared value as low as about 0.64 or as high as about 0.68, with the most likely value of R-squared in the central region of 0.65

Is it possible to predict medical percentage, based on the percentages of specific vaccines that are missing? If so, what are the specifics?

The Frequentist Approach

Build the Prediction Model and Evaluate Model Summary

The dependent variable is the conditional percentages, with the missing vaccines as the independent variables. This model does not include interaction between terms (IVs):

```
predMed <- lm(medical ~ dptMiss + polMiss + mmrMiss + hepMiss + varMiss, data = reportSample)
summary(predMed)
##
## Call:
## lm(formula = medical ~ dptMiss + polMiss + mmrMiss + hepMiss +
##     varMiss, data = reportSample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2650 -0.2099 -0.1660 -0.1495 13.9600
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1495463  0.0457567   3.268  0.00114 **
## dptMiss      0.0008263  0.0132570   0.062  0.95032
## polMiss     -0.0058041  0.0149137  -0.389  0.69726
## mmrMiss      0.0019049  0.0092305   0.206  0.83656
## hepMiss      0.0018727  0.0140864   0.133  0.89427
## varMiss      0.0123548  0.0142314   0.868  0.38562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9374 on 692 degrees of freedom
## Multiple R-squared:  0.01504,    Adjusted R-squared:  0.007922
## F-statistic: 2.113 on 5 and 692 DF,  p-value: 0.06205
```

- The summary output shows the formula applied in the multivariate linear regression model without interaction between terms.
- It then shows the residual summaries, and the intercept together with the B-weights (co-efficient) summaries that desrcribe the line of best fit that reduces the sum of squared errors the most.

- The intercept estimate is 0.1495463 denoting the value of the predicted outcome (medical percentages) when all the missing vaccine values are equal to zero.
- The t-values and the probabilities associated with thoese t-values test the null hypothesis that each of the co-efficients are equal to zero. In this case, we can fail to reject the null hypothesis because of the large p-values for all the independent variables along with an overal model p - value of 0.06205, which is greater than the alpha (p < 0.05)

Re-run the model with interaction between terms

The following model attempts a regression with interaction between terms.

```
predMed <- lm(medical ~ dptMiss * polMiss * mmrMiss * hepMiss * varMiss, data = reportSample)
summary(predMed)
##
## Call:
## lm(formula = medical ~ dptMiss * polMiss * mmrMiss * hepMiss *
##     varMiss, data = reportSample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2007 -0.2447 -0.1027 -0.0149 13.4812
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              1.486e-02  7.322e-02   0.203  0.83929
## dptMiss                                 -2.786e-03  3.863e-02  -0.072  0.94252
## polMiss                                  2.020e-02  3.451e-02   0.585  0.55847
## mmrMiss                                  5.809e-03  1.462e-02   0.397  0.69127
## hepMiss                                  4.288e-03  5.657e-02   0.076  0.93960
## varMiss                                  2.239e-02  6.384e-02   0.351  0.72593
## dptMiss:polMiss                         -2.710e-04  2.459e-03  -0.110  0.91226
## dptMiss:mmrMiss                          1.116e-03  2.951e-03   0.378  0.70549
## polMiss:mmrMiss                         -1.407e-03  3.460e-03  -0.407  0.68431
## dptMiss:hepMiss                          1.471e-02  1.233e-02   1.193  0.23326
## polMiss:hepMiss                         -2.273e-02  1.024e-02  -2.219  0.02680
## mmrMiss:hepMiss                         -5.654e-03  7.653e-03  -0.739  0.46030
## dptMiss:varMiss                         -1.127e-02  1.349e-02  -0.835  0.40382
## polMiss:varMiss                          9.202e-03  1.425e-02   0.646  0.51861
## mmrMiss:varMiss                          1.087e-03  1.001e-02   0.109  0.91352
## hepMiss:varMiss                          1.618e-02  5.700e-03   2.839  0.00467
## dptMiss:polMiss:mmrMiss                 -1.376e-05  4.789e-05  -0.287  0.77399
## dptMiss:polMiss:hepMiss                  1.123e-04  2.834e-04   0.396  0.69209
## dptMiss:mmrMiss:hepMiss                 -3.883e-04  5.348e-04  -0.726  0.46806
## polMiss:mmrMiss:hepMiss                  1.167e-03  5.484e-04   2.128  0.03368
## dptMiss:polMiss:varMiss                  1.054e-04  2.670e-04   0.395  0.69313
## dptMiss:mmrMiss:varMiss                  2.848e-04  5.915e-04   0.481  0.63040
## polMiss:mmrMiss:varMiss                 -5.471e-04  5.833e-04  -0.938  0.34864
## dptMiss:hepMiss:varMiss                 -9.554e-04  7.097e-04  -1.346  0.17870
## polMiss:hepMiss:varMiss                  7.227e-04  7.209e-04   1.002  0.31647
## mmrMiss:hepMiss:varMiss                 -5.633e-04  2.683e-04  -2.099  0.03615
## dptMiss:polMiss:mmrMiss:hepMiss         -1.245e-05  7.348e-06  -1.694  0.09080
## dptMiss:polMiss:mmrMiss:varMiss          4.821e-06  4.766e-06   1.012  0.31208
## dptMiss:polMiss:hepMiss:varMiss         -3.255e-06  4.226e-06  -0.770  0.44145
## dptMiss:mmrMiss:hepMiss:varMiss          4.021e-05  1.810e-05   2.222  0.02664
## polMiss:mmrMiss:hepMiss:varMiss         -2.832e-05  1.758e-05  -1.610  0.10778
## dptMiss:polMiss:mmrMiss:hepMiss:varMiss -4.090e-09  8.702e-09  -0.470  0.63849
##
## (Intercept)
## dptMiss
## polMiss
## mmrMiss
## hepMiss
## varMiss
## dptMiss:polMiss
## dptMiss:mmrMiss
## polMiss:mmrMiss
## dptMiss:hepMiss
## polMiss:hepMiss                         *
## mmrMiss:hepMiss
## dptMiss:varMiss
## polMiss:varMiss
## mmrMiss:varMiss
## hepMiss:varMiss                         **
## dptMiss:polMiss:mmrMiss
## dptMiss:polMiss:hepMiss
## dptMiss:mmrMiss:hepMiss
## polMiss:mmrMiss:hepMiss                 *
## dptMiss:polMiss:varMiss
## dptMiss:mmrMiss:varMiss
## polMiss:mmrMiss:varMiss
## dptMiss:hepMiss:varMiss
## polMiss:hepMiss:varMiss
## mmrMiss:hepMiss:varMiss                 *
## dptMiss:polMiss:mmrMiss:hepMiss         .
## dptMiss:polMiss:mmrMiss:varMiss
## dptMiss:polMiss:hepMiss:varMiss
## dptMiss:mmrMiss:hepMiss:varMiss         *
## polMiss:mmrMiss:hepMiss:varMiss
## dptMiss:polMiss:mmrMiss:hepMiss:varMiss
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9209 on 666 degrees of freedom
## Multiple R-squared:  0.08506,    Adjusted R-squared:  0.04247
## F-statistic: 1.997 on 31 and 666 DF,  p-value: 0.001185
```

In the revised model, there are interactions between terms that are significant, and an overall observed p-value that appears significant. However, further pruning to remove interactions that are not significant is required to improve the model performance.

Additional Model Pruning

Below is another run of the model, with only the significant interactions (p-value < 0.05)

```
predMed <- lm(medical ~ polMiss:hepMiss + dptMiss:polMiss:mmrMiss + polMiss:mmrMiss:hepMiss + mmrMiss:hepMiss:v
arMiss + dptMiss:mmrMiss:hepMiss:varMiss, data = reportSample)

summary(predMed)
```

```
##
## Call:
## lm(formula = medical ~ polMiss:hepMiss + dptMiss:polMiss:mmrMiss +
##     polMiss:mmrMiss:hepMiss + mmrMiss:hepMiss:varMiss + dptMiss:mmrMiss:hepMiss:varMiss,
##     data = reportSample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3093 -0.2032 -0.1424 -0.1316 13.9024
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.316e-01  4.073e-02   3.232 0.001289 **
## polMiss:hepMiss                  1.533e-03  4.210e-04   3.641 0.000292 ***
## polMiss:dptMiss:mmrMiss         -6.841e-07  7.084e-06  -0.097 0.923094
## polMiss:hepMiss:mmrMiss         -2.386e-05  2.632e-05  -0.907 0.364937
## hepMiss:mmrMiss:varMiss         -1.191e-05  1.835e-05  -0.649 0.516576
## hepMiss:dptMiss:mmrMiss:varMiss  2.106e-07  8.385e-08   2.511 0.012267 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9332 on 692 degrees of freedom
## Multiple R-squared:  0.02385,    Adjusted R-squared:  0.0168
## F-statistic: 3.382 on 5 and 692 DF,  p-value: 0.005019
```

The resulting model still bears interactions that are not significant. Further pruning yields the outcome below:

```
predMed <- lm(medical ~ polMiss:hepMiss + dptMiss:mmrMiss:hepMiss:varMiss, data = reportSample)
summary(predMed)
##
## Call:
## lm(formula = medical ~ polMiss:hepMiss + dptMiss:mmrMiss:hepMiss:varMiss,
##     data = reportSample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7493 -0.1949 -0.1777 -0.1752 14.0339
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.752e-01  3.796e-02   4.614 4.69e-06 ***
## polMiss:hepMiss                  3.182e-04  1.192e-04   2.669  0.00778 **
## hepMiss:dptMiss:mmrMiss:varMiss -4.310e-08  1.704e-08  -2.530  0.01163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9376 on 695 degrees of freedom
## Multiple R-squared:  0.01028,    Adjusted R-squared:  0.00743
## F-statistic: 3.609 on 2 and 695 DF,  p-value: 0.0276
```

Finally, all p-values of the coefficients are less than alpha threshold $p < 0.05$.

The model summary shows the multiple R-squared value of 0.01028 which shows very littel of the variation in medical percentages can be accounted for by the predictor missing vaccine variables working interactively.

The Null hypothesis Significance test (F-test) is whether the R-squared value is significantly different from zero. In this case we reject the null hypothesis, because the model p-value of < 2e-16 is much less than the alpha threshold of $p < 0.05$. The test is statistically significant.

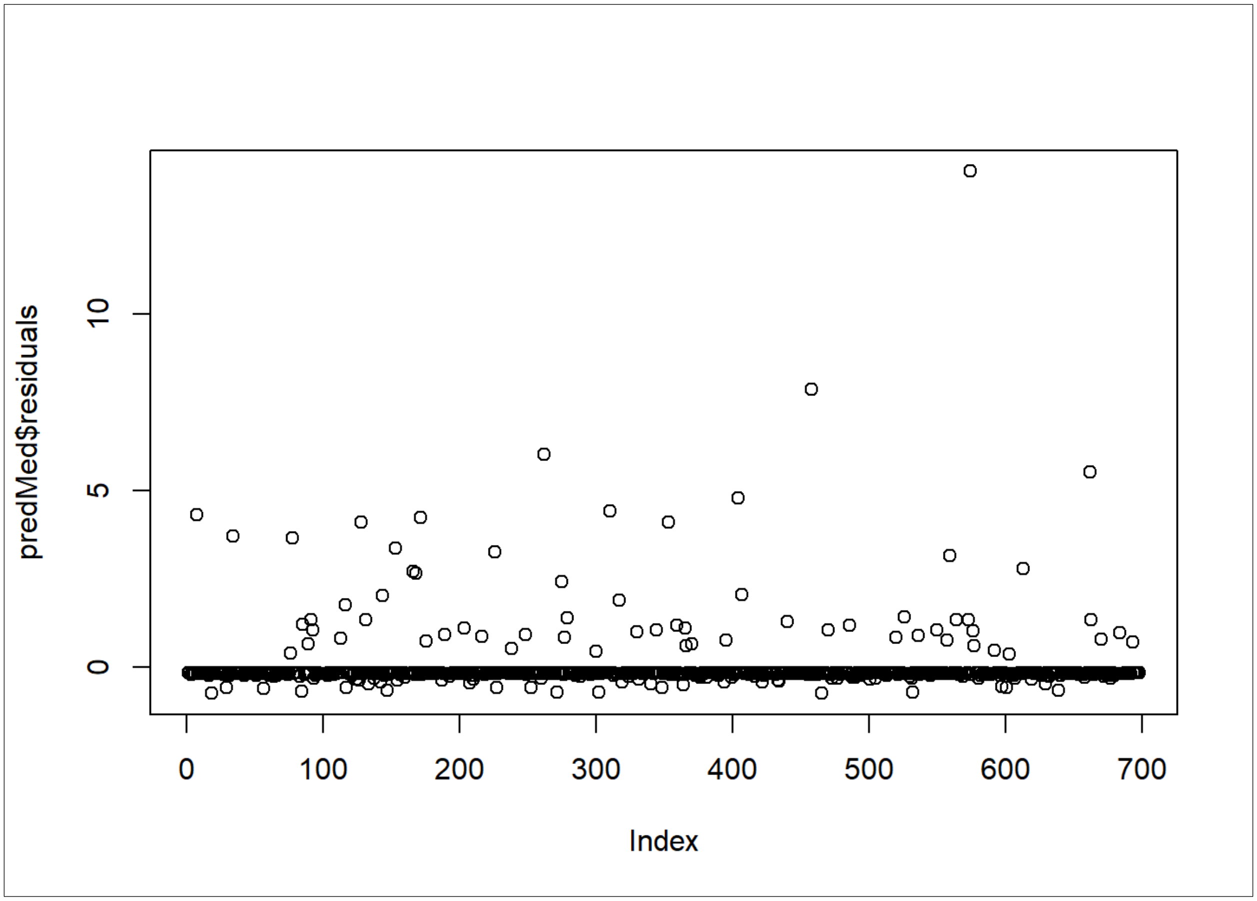
Thus, the prediction of medical percentages based on missing vaccines is as follows, where the bracket denotes multiplication:

Medical rates = 1.752e-01 + 3.182e-04(polMiss:hepMiss) + -4.310e-08(hepMiss:dptMiss:mmrMiss:varMiss)

Analyzing Residuals for normality/non-normality

The mean of the residuals from the least squares processing should always be zero.

```
plot(predMed$residuals)
```



The plot shows a largely normal distribution with the mean centered around zero indicating that the underlying relationship between the independent variables and the dependent variables are linear.

Conclusion:

Even though the model has a statistically significant p - value, the R-Squared value of 0.01028 (adjusted to 0.00743) means very little of the variability within the dependent variable can be explained by interactions between the independent variables. Thus, the prediction of medical percentages cannot be meaningfully done with the missing vaccine variables.

Bayesian Approach

Build the Prediction Model and Evaluate Model Summary

```
regOutMCMC <- lmBF(medical ~ polMiss:hepMiss + dptMiss:mmrMiss:hepMiss:varMiss, data = reportSample, posterior=
TRUE, iterations = 10000)


summary(regOutMCMC)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                             Mean        SD  Naive SE
## mu                                     2.089e-01 3.556e-02 3.556e-04
## polMiss.&.hepMiss                      3.087e-04 1.170e-04 1.170e-06
## hepMiss.&.dptMiss.&.mmrMiss.&.varMiss -4.189e-08 1.667e-08 1.667e-10
## sig2                                   8.801e-01 4.708e-02 4.708e-04
## g                                      1.339e-01 4.897e-01 4.897e-03
##                                       Time-series SE
## mu                                         3.556e-04
## polMiss.&.hepMiss                          1.170e-06
## hepMiss.&.dptMiss.&.mmrMiss.&.varMiss      1.667e-10
## sig2                                       4.791e-04
## g                                          4.897e-03
##
## 2. Quantiles for each variable:
##
##                                             2.5%        25%        50%
## mu                                     1.394e-01  1.844e-01  2.086e-01
## polMiss.&.hepMiss                      7.800e-05  2.299e-04  3.076e-04
## hepMiss.&.dptMiss.&.mmrMiss.&.varMiss -7.483e-08 -5.305e-08 -4.177e-08
## sig2                                   7.916e-01  8.477e-01  8.782e-01
## g                                      1.430e-02  3.295e-02  5.774e-02
##                                              75%      97.5%
## mu                                     2.330e-01  2.785e-01
## polMiss.&.hepMiss                      3.884e-04  5.345e-04
## hepMiss.&.dptMiss.&.mmrMiss.&.varMiss -3.087e-08 -8.862e-09
## sig2                                   9.107e-01  9.764e-01
## g                                1.146e-01   6.752e-01
```

The output shows the means of the respective distributions and the 95% HDIs. The mean column is the paremeter estimates of the B-Weights of each of the predictions and they are a very close match to the frequentist calculations prior.

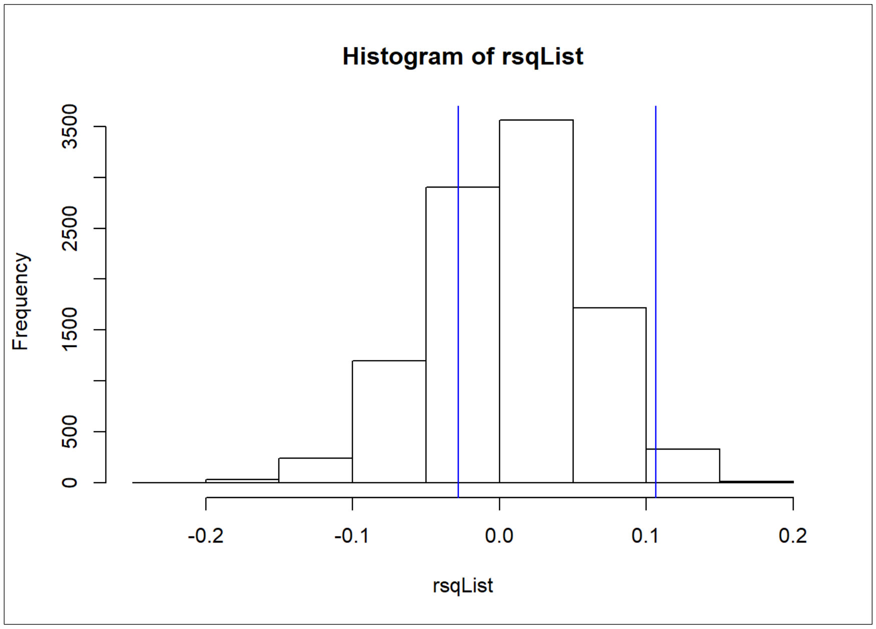
The 2.5% and 97.5% boundaries of 95% HDIs for the IVs are very similar exept for hepMiss - this reflects similar finding in the frequentist approach where hepMiss was dropped from the initial model.

The sig2(sigma-squared) shows the model precision for each of the iterations. The smaller sig2 is, the better the prediction.

Converting Sigma-squared values to R-squared values

The R-squared value of each model in the posterior distribution is equal to 1 minus the value of sig2 divided by the variance in the dependent variable.

```
rsqList <- 1 - (regOutMCMC[,"sig2"] / var(reportSample$medical))
# mean
mean(rsqList)
## [1] 0.006354223
hist(rsqList)
# lower bound of 95% HDI
abline(v=quantile(rsqList, c(0.25)), col="blue")
# upper bound of 95% HDI
abline(v=quantile(rsqList, c(0.975)), col="blue")
```



Conclusions

The mean value of the distribution came to 0.007 which is exactly similar to the adjusted R-squared value obtained using the frequentist approach prior. The Bayesian model presents visual look at the likely range of possibilities for the predictive strength of the model. It is possible to expect an R-squared value as low as about -0.002 or as high as about 0.011, with the most likely value of R-squared in the central region of 0.00.

The Bayesian outcome confirms the frequentist outcome that it is not possible to accurately predict medical exemption rates based on the missing vaccine variables.

Is it possible to predict religious percentage, based on the percentages of specific vaccines that are missing? If so, what are the specifics?

Frequentist Approach

Build the Prediction Model and Evaluate Model Summary

The dependent variable is the religious exemption percentages, with the missing vaccines as the independent variables.

```
predRel <- lm(religious ~ dptMiss + polMiss + mmrMiss + hepMiss + varMiss, data = reportSample)
summary(predRel)
```

```
## 
## Call:
## lm(formula = religious ~ dptMiss + polMiss + mmrMiss + hepMiss + 
##     varMiss, data = reportSample)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -71.474  -0.553  -0.024   0.938  22.499 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.27002    0.22094   1.222  0.22208    
## dptMiss     -0.17365    0.06401  -2.713  0.00684 ** 
## polMiss     -0.01645    0.07201  -0.228  0.81941    
## mmrMiss      0.05361    0.04457   1.203  0.22949    
## hepMiss      0.28468    0.06802   4.185 3.21e-05 ***
## varMiss      0.56384    0.06872   8.205 1.12e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.526 on 692 degrees of freedom
## Multiple R-squared:  0.7263, Adjusted R-squared:  0.7243 
## F-statistic: 367.3 on 5 and 692 DF,  p-value: < 2.2e-16
```

Initial Pruning

Drop polMiss and mmrMiss from the model and run:

```
predRel <- lm(religious ~ dptMiss + hepMiss + varMiss, data = reportSample)
summary(predRel)
##
## Call:
## lm(formula = religious ~ dptMiss + hepMiss + varMiss, data = reportSample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.606  -0.529  -0.018   0.954  22.390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.29353    0.22000   1.334    0.183
## dptMiss     -0.14685    0.02845  -5.162 3.20e-07 ***
## hepMiss      0.28496    0.06694   4.257 2.36e-05 ***
## varMiss      0.57501    0.06805   8.450  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.525 on 694 degrees of freedom
## Multiple R-squared:  0.7257, Adjusted R-squared:  0.7246
## F-statistic: 612.1 on 3 and 694 DF,  p-value: < 2.2e-16
```

Finally, all p-values of the coefficients are less than alpha threshold $p < 0.05$.

The model summary shows the multiple R-squared value of 0.7257 which shows significant proportion of the variations in religious percentages can be accounted for by the predictor missing vaccine variables working interactively.

The Null hypothesis Significance test (F-test) is whether the R-squared value is significantly different from zero. In this case we reject the null hypothesis, because the model p-value of < 2.2e-16 is much less than the alpha threshold of $p < 0.05$. The test is statistically significant.

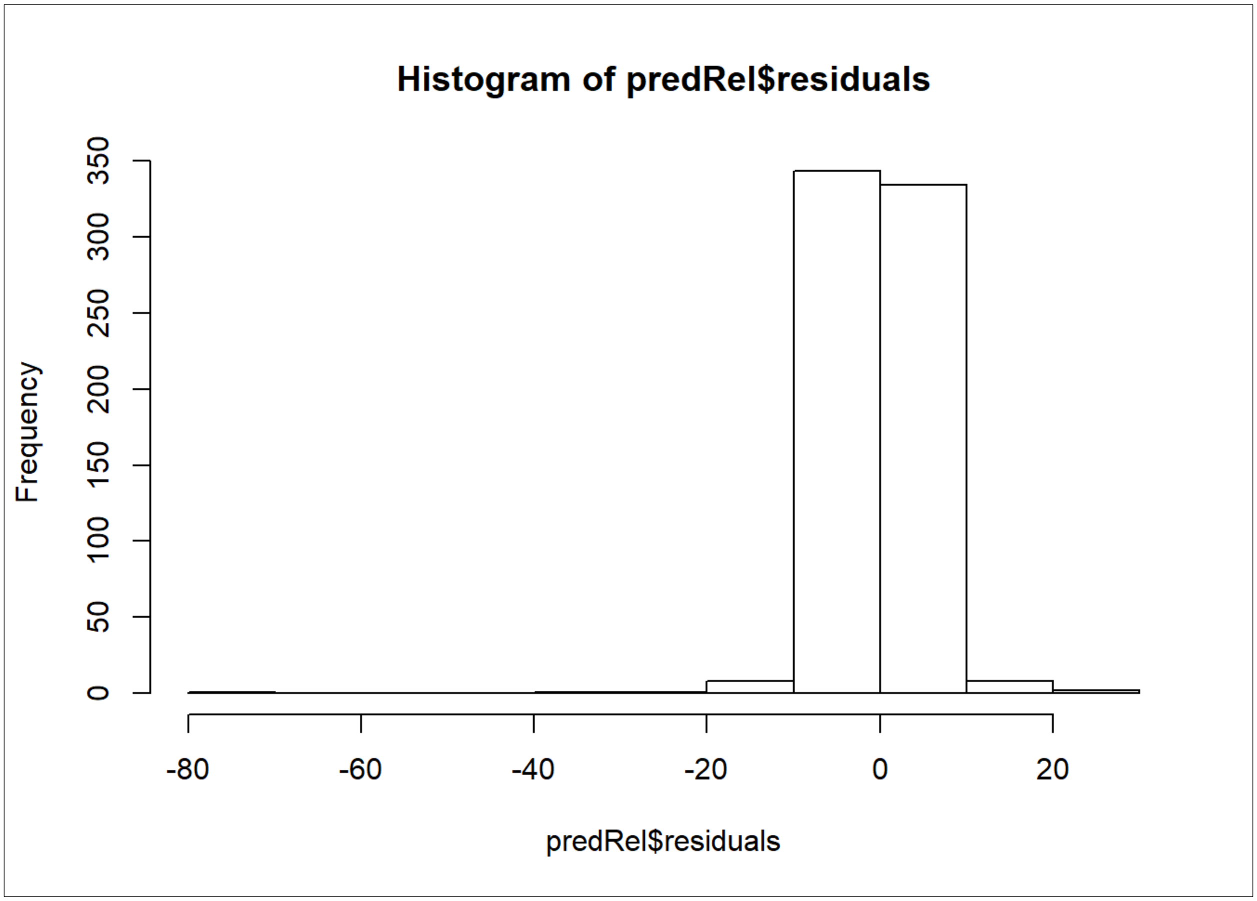
Thus, the prediction of religious exemption percentages based on missing vaccines is as follows, where the bracket denote multiplication:

religious = 0.29353 - 0.14685(dptMiss) + 0.28496(hepMiss) + 0.57501(varMiss)

Analyzing Residuals for normality/non-normality

The mean of the residuals from the least squares processing should always be zero.

```
hist(predRel$residuals)
```



The plot shows a largely normal distribution with the mean centered around zero indicating that the underlying relationship between the independent variables and the dependent variables are linear.

Conclusion:

Since the model has a statistically significant p - value of < 2.2e-16, the R-Squared value of 0.7257 (adjusted to 0.7246) means significant proportion of the variability within the dependent variable can be explained by the independent variables working together. Thus, the prediction of religious exemption percentages can be meaningfully done with the missing vaccine variables as follows:

religious = 0.29353 - 0.14685(dptMiss) + 0.28496(hepMiss) + 0.57501(varMiss)

Bayesian Approach

Build the Prediction Model and Evaluate Model Summary

```
regOutMCMC <- lmBF(religious ~ dptMiss + hepMiss + varMiss, data = reportSample, posterior=TRUE,
iterations = 10000)


summary(regOutMCMC)
```

```
##

## Iterations = 1:10000

## Thinning interval = 1

## Number of chains = 1

## Sample size per chain = 10000

##

## 1. Empirical mean and standard deviation for each variable,

##    plus standard error of the mean:

##

##             Mean      SD  Naive SE Time-series SE

## mu       4.2140 0.17320 0.0017320      0.0017955

## dptMiss -0.1466 0.02858 0.0002858      0.0002859

## hepMiss  0.2848 0.06686 0.0006686      0.0006682

## varMiss  0.5736 0.06795 0.0006795      0.0006805

## sig2    20.5447 1.11970 0.0111970      0.0111970

## g        1.3529 2.64439 0.0264439      0.0271768

##

## 2. Quantiles for each variable:

##

##             2.5%     25%     50%     75%    97.5%

## mu       3.8811  4.0949  4.2150  4.3305  4.55434

## dptMiss -0.2026 -0.1661 -0.1467 -0.1272 -0.09044

## hepMiss  0.1540  0.2393  0.2846  0.3303  0.41378

## varMiss  0.4381  0.5276  0.5738  0.6190  0.70518

## sig2    18.5074 19.7693 20.4965 21.2751 22.88149

## g        0.2398  0.4987  0.8158  1.4262  5.66111
```

The output shows the means of the respective distributions and the 95% HDIs. The mean column is the paremeter estimates of the B-Weights of each of the predictions and they are a very close match to the frequentist calculations prior.

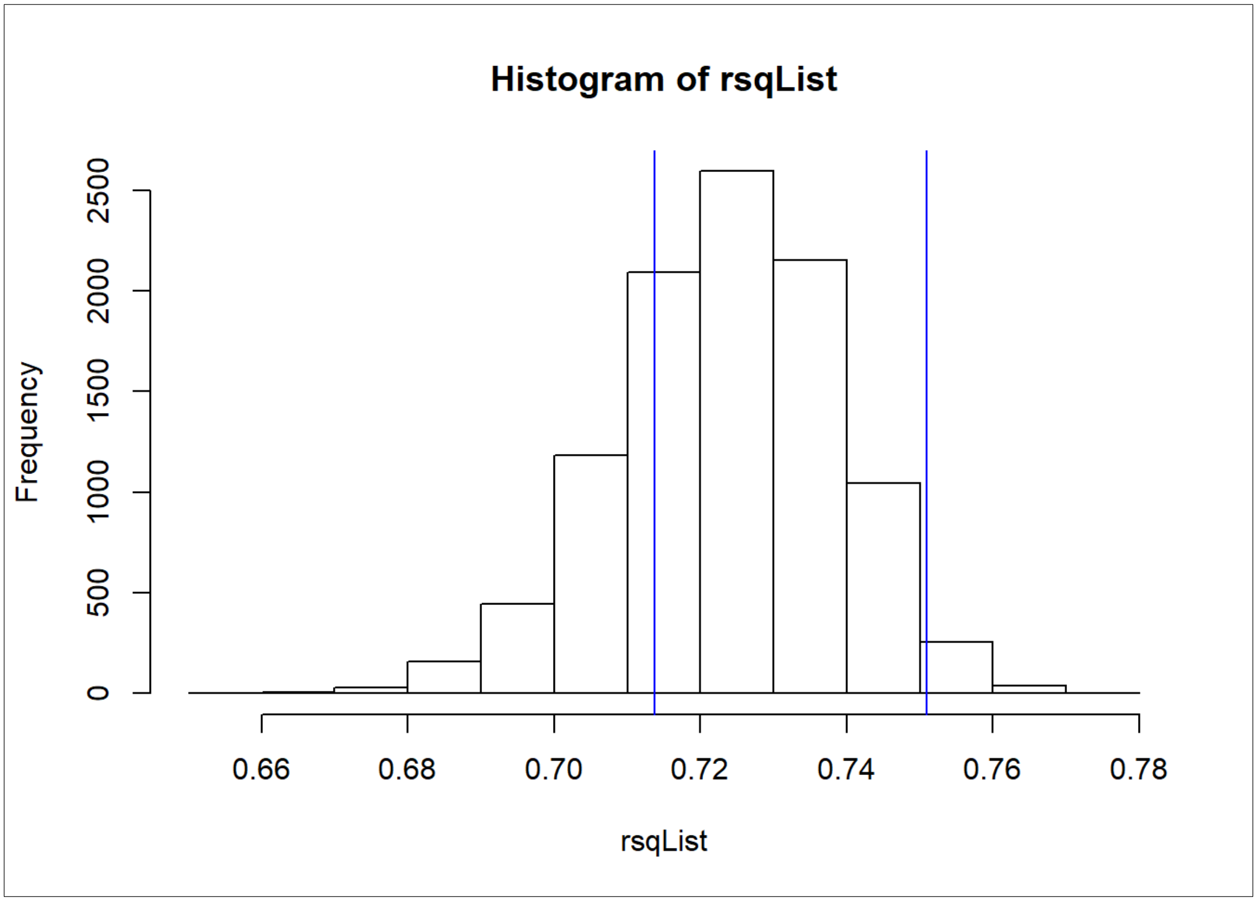
The 2.5% and 97.5% boundaries of 95% HDIs for the IVs are very similar exept for hepMiss - this reflects similar finding in the frequentist approach where hepMiss was dropped from the initial model.

The sig2(sigma-squared) shows the model precision for each of the iterations. The smaller sig2 is, the better the prediction.

Converting Sigma-squared values to R-squared values

The R-squared value of each model in the posterior distribution is equal to 1 minus the value of sig2 divided by the variance in the dependent variable.

```
rsqList <- 1 - (regOutMCMC[,"sig2"] / var(reportSample$religious))
# mean
mean(rsqList)
## [1] 0.7235743
hist(rsqList)
# lower bound of 95% HDI
abline(v=quantile(rsqList, c(0.25)), col="blue")
# upper bound of 95% HDI
abline(v=quantile(rsqList, c(0.975)), col="blue")
```



Conclusions

The mean value of the distribution came to 0.7233177 which is very similar to the adjusted R-squared value obtained using the frequentist approach prior. The Bayesian model presents visual look at the likely range of possibilities for the predictive strength of the model. It is possible to expect an R-squared value as low as about 0.71 or as high as about 0.75, with the most likely value of R-squared in the central region of 0.72.

The Bayesian outcome confirms the frequentist outcome that it is possible to accurately predict medical exemption rates based on the missing vaccine variables with credible model.

What's the big picture, based on all the foregoing analyses?

Below are some of the big picture items gleaned from this analysis:

- Public kindergartens are nearly four times as many as private ones in California.
- There are 3 private kindergartens that don't report vaccines for every public kindergarten that doesn't.
- It is inconclusive whether a prediction of school type is possible based on conditional, religious or medical exemption rates alone. The frequentist prediction, when attempted, is not reproducible using Bayesian model hence its viability is inconclusive
- Religious exemption rates can be predicted with greater accuracy from the missing vaccine records.
- However, it is not possible to predict the medical exemption rates based solely on the relationships between other missing vaccine reports
- All Vaccine exemption rates are very different in public vs private kindergartens indicating that there are other reasons outside of the dataset, perhaps, driving these differences. Possible reasons maybe socio-economic data because it is generalizable that private kindergartens are more expensive, hence mostly frequented by higher socio-economic status children.
- The U.S polio vaccination trend has reached its peak and plateaued. This is either good news or bad news, depending on the future number of new cases of polio recorded.

REFERENCES

History of Vaccines, (2020). Retrieved from https://www.historyofvaccines.org/timeline#EVT_100348. Retrieved Tuesday, March 31, 2020.