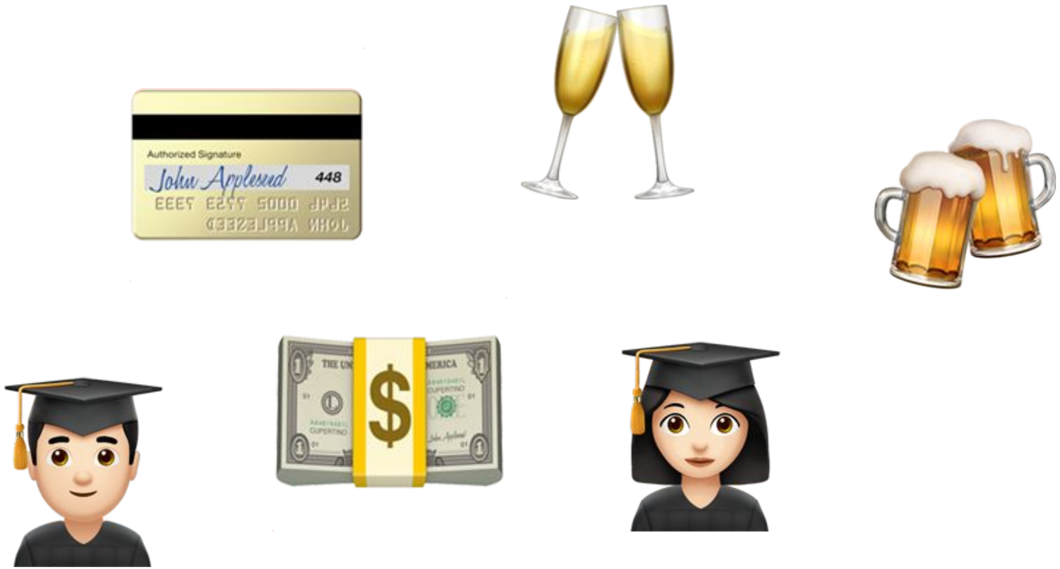
Predict the Overdraft among College Students from the perspective of a Frequentist and a Bayesian



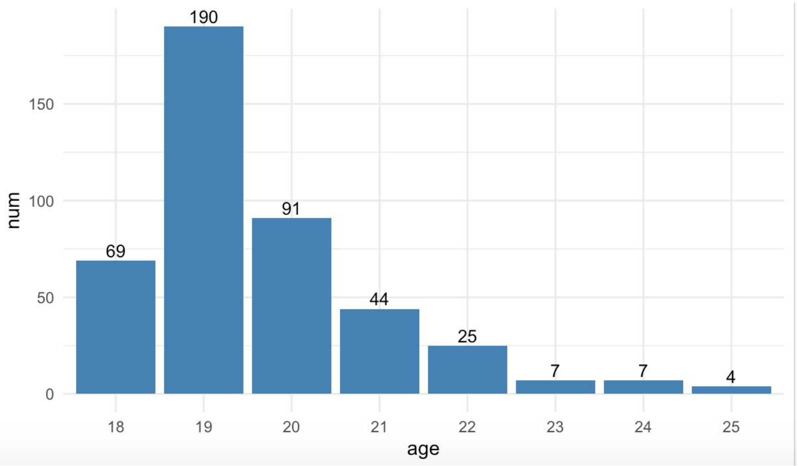
Yunjian Hao

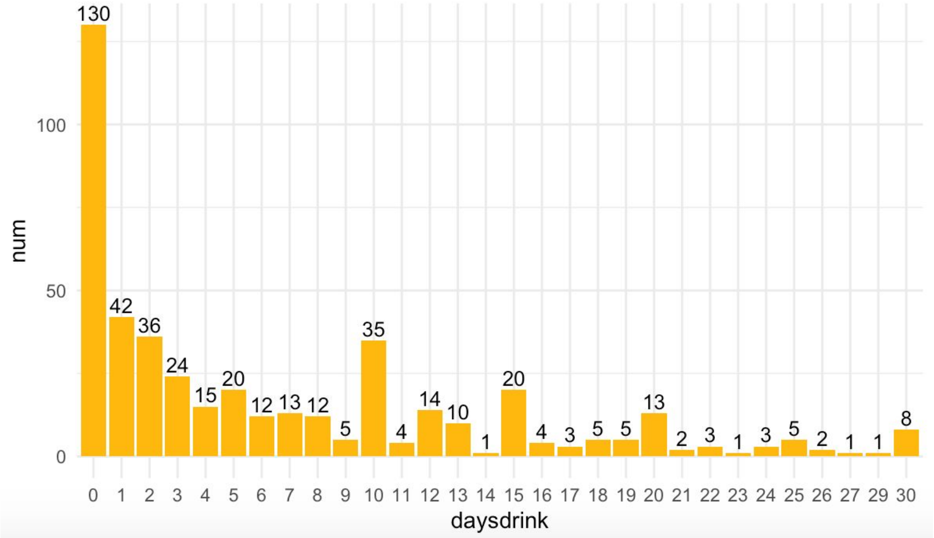
Mengchen Xu

Tingjue Yin

Dataset Description

- A survey of 450 undergraduates in the universities in Mississippi
- Factors (Age, Sex, Daysdrink) related to having overdrawn a checking account
- Sex (0: 197 males, 1: 252 females)





Frequentist Perspective: modeling

```
> logit_reg_freq <- glm(Overdrawn ~ Age + Sex + DaysDrink, data=df_train, family=binomial(link="logit"))
> summary(logit_reg_freq)

Call:
glm(formula = Overdrawn ~ Age + Sex + DaysDrink, family = binomial(link = "logit"),
    data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5975  -0.5886  -0.4784  -0.3212   2.5021

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.46578    2.10211  -4.027 5.64e-05 ***
Age          0.27635    0.10097   2.737 0.006200 **
Sex          1.16097    0.36669   3.166 0.001545 **
DaysDrink    0.06484    0.01967   3.297 0.000977 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 290.64  on 349  degrees of freedom
Residual deviance: 267.38  on 346  degrees of freedom
AIC: 275.38

Number of Fisher Scoring iterations: 5
```

Frequentist Perspective: Predict & Result

```
> predicted_response <- plogis(predict(logit_reg_freq, df_test))  # predicted scores
> cutoff <- optimalCutoff(df_test, predicted_response)[1]
> cutoff
[1] 0.1071024
> data1 = as.numeric(predicted_response>cutoff)
> confusionMatrix(data = as.factor(data1), reference = as.factor(df_test$Overdrawn))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 47  3
         1 29  8

               Accuracy : 0.6322
                 95% CI : (0.522, 0.7331)
    No Information Rate : 0.8736
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1719

 Mcnemar's Test P-Value : 9.897e-06

            Sensitivity : 0.6184
            Specificity : 0.7273
         Pos Pred Value : 0.9400
         Neg Pred Value : 0.2162
             Prevalence : 0.8736
         Detection Rate : 0.5402
   Detection Prevalence : 0.5747
      Balanced Accuracy : 0.6728

       'Positive' Class : 0

> |
```

Bayesian Perspective

```
log_reg_model <- "
  data{
  int <lower = 0> n;
  int <lower=0, upper = 1> Y[n];
  vector[n] X1;
  vector[n] X2;
  vector[n] X3;
  }

  parameters{
    real beta_0;
    real beta_1;
    real beta_2;
    real beta_3;
  }

model{
  Y ~ bernoulli_logit(beta_0 + beta_1*X1 + beta_2*X2 + beta_3*X3);
  beta_1 ~ normal(0.075,10);
  beta_2 ~ normal(0.295,10);
  beta_3 ~ normal(0.1,10);
  }
"
```

Bayesian Perspective: Prior

	Age	Sex	DaysDrink	Overdrawn
1	19	0	20	0
2	19	1	7	0
3	19	0	5	0

likelihood:

Y ~ bernoulli_logit(beta_0 + beta_1*X1 + beta_2*X2 + beta_3*X3)

β0

The log odds of overdraw a checking account when all features equal to 0, which means the male student at an age of 0 doesn't drink over the past month the log odds of overdraw a checking account. Hence β0 has no meaningful interpretation in our situation.

β1 (Age) Beta_1 ~ normal(0.075,10)

- Students' financial behavior scores were significantly related to age (*p* = .015) and gender (*p* < .001).
- Older students tended to have a higher number of problem financial behaviors. Each additional year of age was associated with a 7.5% increase in the average number of problem financial behaviors.

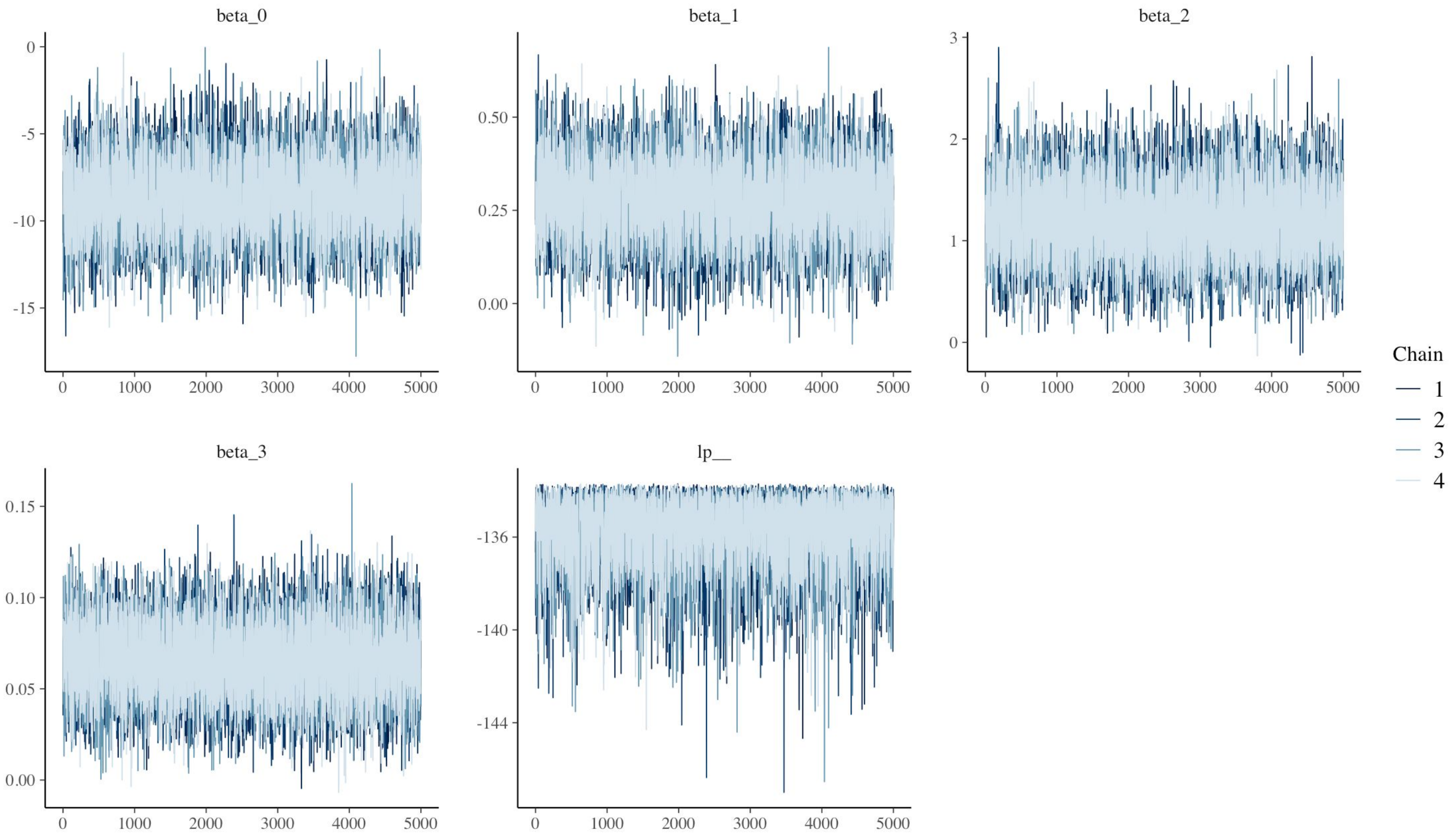
β2 (Sex) Beta_2 ~ normal(0.295,10)

- Female students tended to have more problematic financial behaviors than male students. Male students had approximately 29.5% fewer problem financial behaviors compared to female students. (Adams&Moore, 2007)

β3 (DaysDrink) Beta_3 ~ normal(0.1,10)

- 10 out of 15 drunk and no one had the overdraft but 2 of them reported out of money

Bayesian Perspective: Plot distribution



Bayesian Perspective: Predict rstan:

```
newdf <- as.array(log_reg_sim) %>%
  reshape2::melt() %>%
  pivot_wider(names_from = parameters,
              values_from = value)

# get mode func
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

predict = c()
for (i in 1:nrow(df_test)){
  prob_array = c()
  row <- df_test[i,]
  logi = newdf$beta_0 + newdf$beta_1*row[1,1] +
    newdf$beta_2*row[1,2] + newdf$beta_3*row[1,3]
  prob = (exp(1)^logi)/(1+(exp(1)^logi))
  p = rbinom(1,size=1, prob=prob)
  prob_array = c(prob_array, p)
  mode = getmode(prob_array)
  predict = c(predict,mode)
}
```

df_test

	Age	Sex	DaysDrink	Overdrawn
1	19	0	20	0
2	19	1	7	0
3	19	0	5	0
4	19	1	0	0
5	19	1	0	0

newdf

	iterations	chains	beta_0	beta_1	beta_2	beta_3	lp__
1	1	chain:1	-7.171407	0.22682842	0.69816790	0.08739458	-136.6545
2	2	chain:1	-7.154966	0.22527407	0.69709346	0.08587830	-136.3092
3	3	chain:1	-10.784473	0.38190952	1.35444714	0.06985489	-134.3178
4	4	chain:1	-10.596316	0.37790829	1.04908400	0.08219555	-134.8376
5	5	chain:1	-10.362507	0.35367465	1.40102763	0.07840622	-134.3313
6	6	chain:1	-9.359735	0.32791336	0.96197017	0.06149714	-134.0625
7	7	chain:1	-9.388161	0.31546335	1.17149402	0.06993822	-133.9716
8	8	chain:1	-9.078983	0.30320809	1.23535770	0.05332059	-134.2458
9	9	chain:1	-10.773301	0.39378169	1.08284822	0.05975892	-134.5783
10	10	chain:1	-11.400647	0.42207402	1.14940467	0.07327690	-134.9095

Bayesian Perspective: rstan Result

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 65  6
         1 14  2

               Accuracy : 0.7701
                 95% CI : (0.6675, 0.8536)
    No Information Rate : 0.908
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0502

 Mcnemar's Test P-Value : 0.1175

            Sensitivity : 0.8228
            Specificity : 0.2500
         Pos Pred Value : 0.9155
         Neg Pred Value : 0.1250
             Prevalence : 0.9080
         Detection Rate : 0.7471
   Detection Prevalence : 0.8161
      Balanced Accuracy : 0.5364

       'Positive' Class : 0
```

Bayesian Perspective: Predict rstanarm:

```
log_reg_arm <- stan_glm(Overdrawn ~ Age+Sex+DaysDrink,
                        data = df_train,
                        family = binomial(link = 'logit'))

pred_test_arm <- posterior_predict(log_reg_arm,
                                   newdata = df_test)
pred_test_arm

predict1 = c()
for (i in 1:nrow(df_test)){
  col <- pred_test_arm[,i]
  mode = getmode(as.vector(col))
  predict1 = c(predict1,mode)
}


confusionMatrix(data = as.factor(predict1), reference = as.factor(df_test$Overdrawn))
```

Bayesian Perspective: rstanarm Result

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 78  9
         1  0  0

               Accuracy : 0.8966
                 95% CI : (0.8127, 0.9516)
    No Information Rate : 0.8966
    P-Value [Acc > NIR] : 0.587539

                  Kappa : 0

 Mcnemar's Test P-Value : 0.007661

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8966
         Neg Pred Value :    NaN
             Prevalence : 0.8966
         Detection Rate : 0.8966
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

Reference

Worthy S.L., Jonkman J.N., Blinn-Pike L. (2010), "Sensation-Seeking, Risk-Taking, and Problematic Financial Behaviors of College Students," Journal of Family and Economic Issues, 31: 161-170

Adams, T., & Moore, M. (2007). High-risk health and credit behavior among 18- to 25-year-old college students. *Journal of American College Health, 56*, 101–108.