Midterm Exam

Zhaosheng Xie

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

Background: In my undergraduate university' dining hall, the only one way for students to pay their food was to use student card. But dining hall published a new way last year, which is to use cellphone. Students can bind their student ID to wechat wallet and use wechat payment.

Data description: In my data, I observe 4 groups, which are 4 food windows. For each window, I observe 10 students whose attributes are divided into 3. The first is cad/phone column. 0 represents student card and 1 represents cellphone. In gender column, 0 represents female and 1 represents male. Last is backpack column. 0 represents students do not have a backpack when they order food and 1 is otherwise. In this case, I consider that wearing backpack means students just finished their class and went for a lunch. Without backpack means students did not have classes. Why this is important? Because students must use their ID card to sign in at classroom. In conclusion, students who just finished classes must carry ID card and other students' condition is not for sure. I assume every student carries their mobile phone.

Comparison of interest: 1. I want to find which way students prefer. 2. Whether genders and wearing backpack affect students' decision of pattern of payment.

```
library(readr)
Data <- read_csv("Data from studentes.csv")
```

```
## Parsed with column specification:
## cols(
```

```
##   `student id` = col_double(),
##   `window name` = col_character(),
##   `card/phone` = col_double(),
##   gender = col_double(),
##   backpack = col_double()
## )
```

```
head(Data, n = 10)
```

```
## # A tibble: 10 x 5
##    `student id` `window name`  `card/phone` gender backpack
##           <dbl> <chr>                 <dbl>  <dbl>    <dbl>
##  1            1 Orient Express            1      1        0
##  2            2 Orient Express            1      1        0
##  3            3 Orient Express            1      1        1
##  4            4 Orient Express            1      1        0
##  5            5 Orient Express            1      1        1
##  6            6 Orient Express            1      1        1
##  7            7 Orient Express            1      1        1
##  8            8 Orient Express            1      1        0
##  9            9 Orient Express            0      1        0
## 10           10 Orient Express            0      0        0
```

EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```
## data cleaning
# change "1" and "0" into character
Data1 <- Data # to keep original data
for (i in 1:40){
  Data1$`card/phone`[i] <- ifelse(Data1$`card/phone`[i]==1, "phone", "card")
  Data1$gender[i] <- ifelse(Data1$gender[i]==1, "M", "F")
  Data1$backpack[i] <- ifelse(Data1$backpack[i]==1, "Wear", "Not wear")
}
# rename columns
colnames(Data1)[2] <- "window"
colnames(Data1)[3] <- "pay_way"
head(Data1)
```

```
## # A tibble: 6 x 5
##   `student id` window         pay_way gender backpack
##          <dbl> <chr>          <chr>   <chr>  <chr>
## 1            1 Orient Express phone   M      Not wear
## 2            2 Orient Express phone   M      Not wear
## 3            3 Orient Express phone   M      Wear
## 4            4 Orient Express phone   M      Not wear
## 5            5 Orient Express phone   M      Wear
## 6            6 Orient Express phone   M      Wear
```

```
Data2 <- Data
colnames(Data2)[2] <- "window"
colnames(Data2)[3] <- "pay_way"
```

```r
# pay way frequency
library(funModeling)
```

```
## Warning: package 'funModeling' was built under R version 4.0.3
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

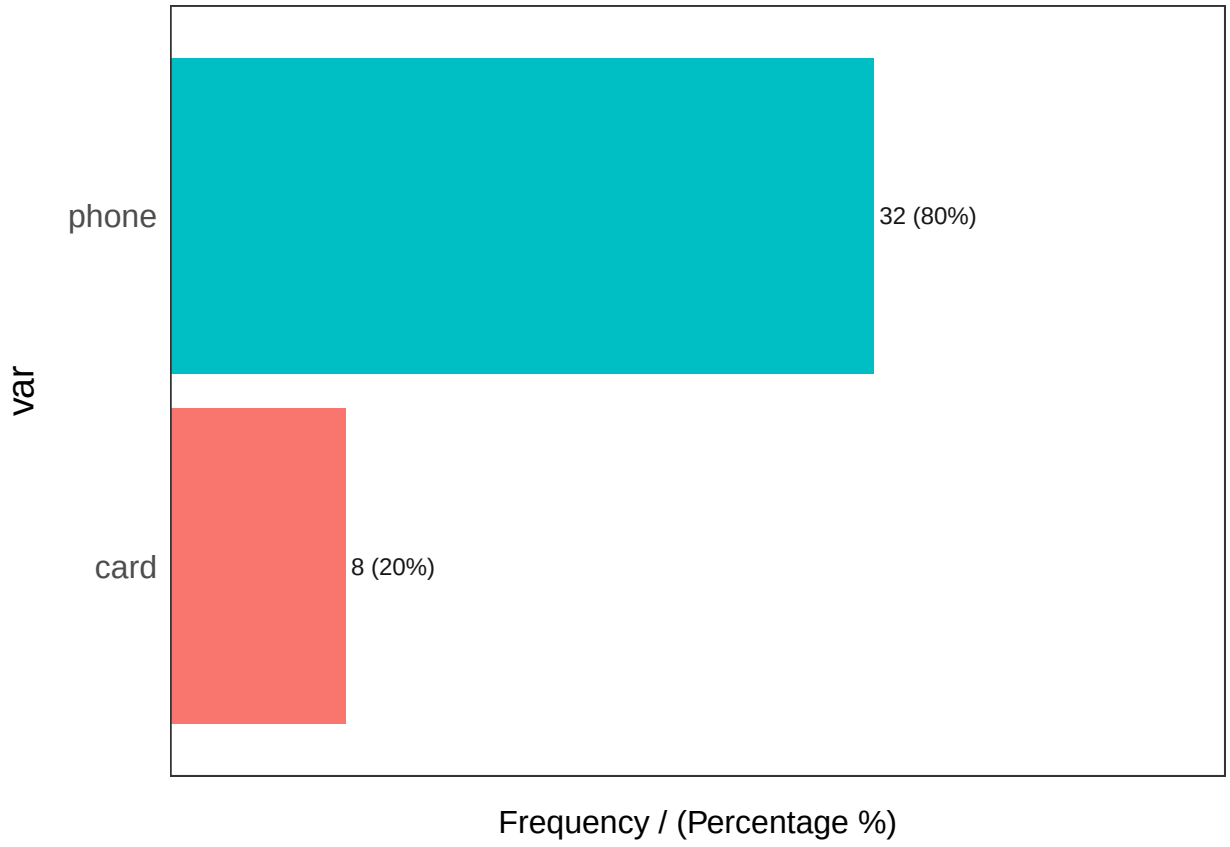
```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## funModeling v.1.9.4 :)
## Examples and tutorials at livebook.datascienceheroes.com
##  / Now in Spanish: librovivodecienciadedatos.ai
```

```r
Pr.pay_way <- freq(Data1$pay_way)
```




```r
Pr.pay_way
```

```
##     var frequency percentage cumulative_perc
```

```
## 1 phone        32         80              80
## 2  card         8         20             100
```

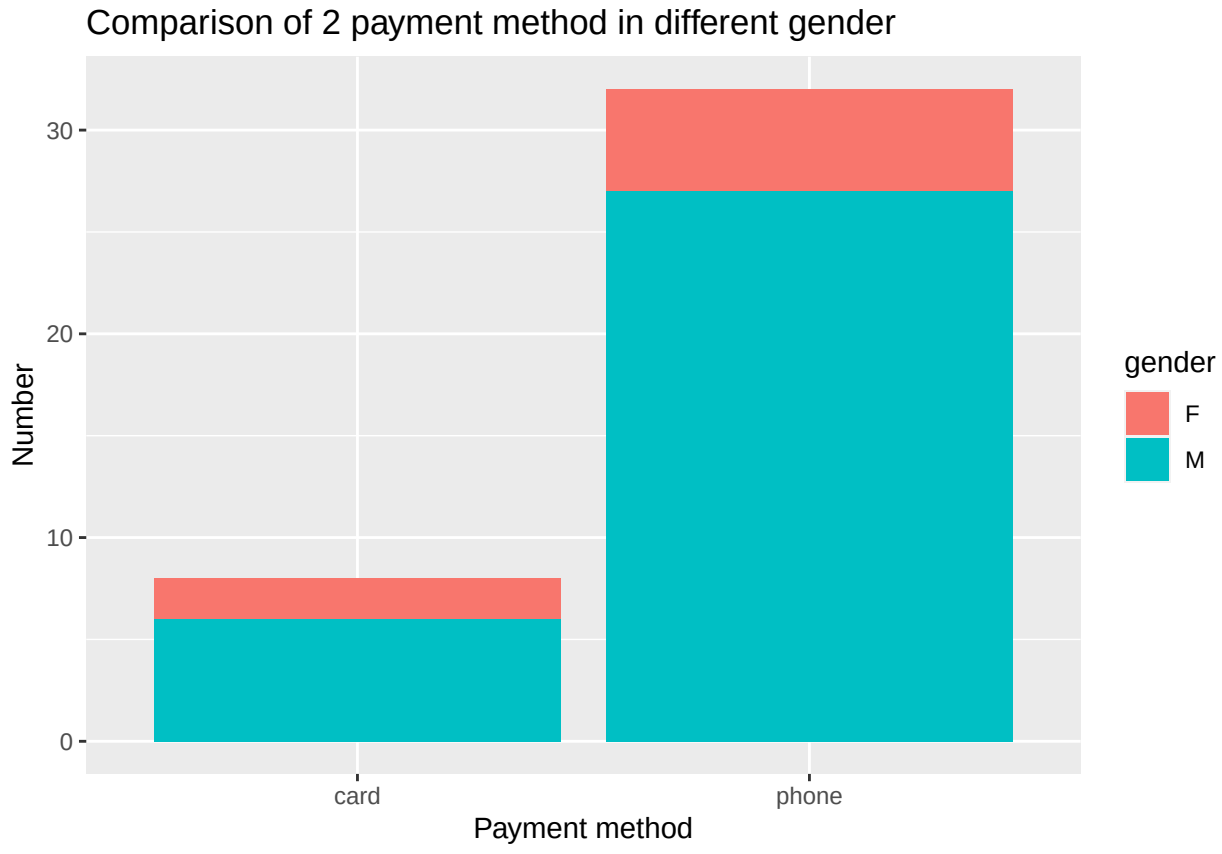
```
## Comparison of 2 payment ways
# different gender
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidy
```

```
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v purrr   0.3.4     v forcats 0.5.0
```

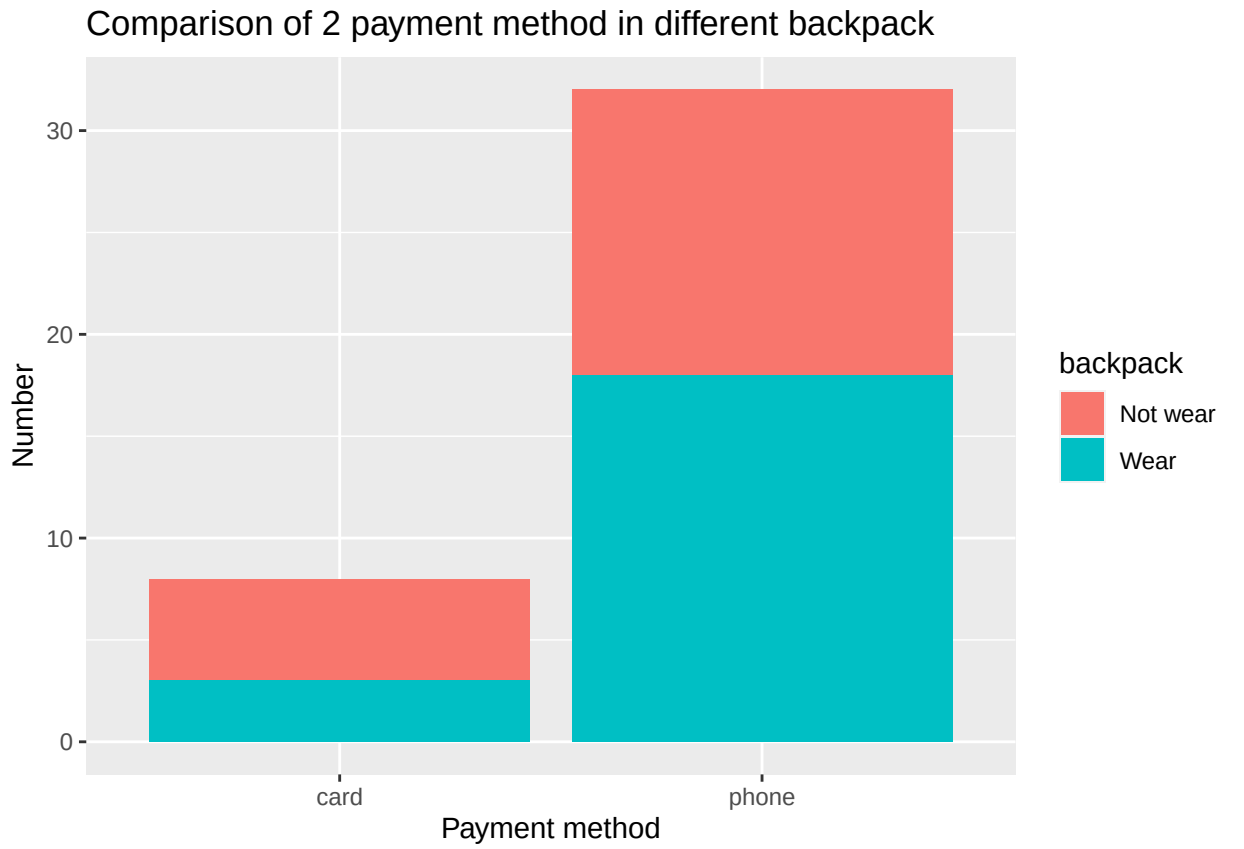
```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x dplyr::src()       masks Hmisc::src()
## x dplyr::summarize() masks Hmisc::summarize()
```

```
library(ggplot2)
ggplot(data = Data1) +
  geom_bar(mapping = aes(x = pay_way, fill = gender)) +
  labs(x = "Payment method", y = "Number",
       title = "Comparison of 2 payment method in different gender")
```

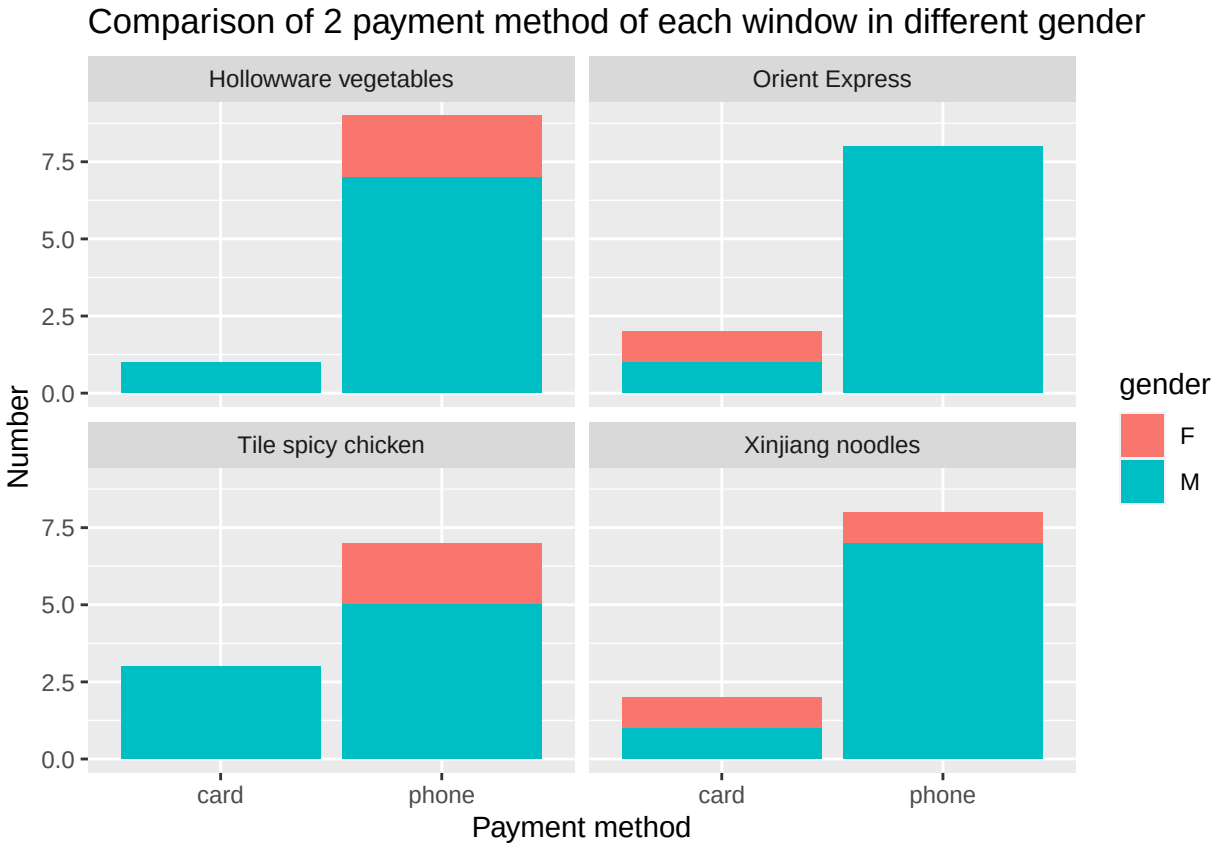



```
# different backpack condition
ggplot(data = Data1) +
  geom_bar(mapping = aes(x = pay_way, fill = backpack)) +
  labs(x = "Payment method", y = "Number",
```

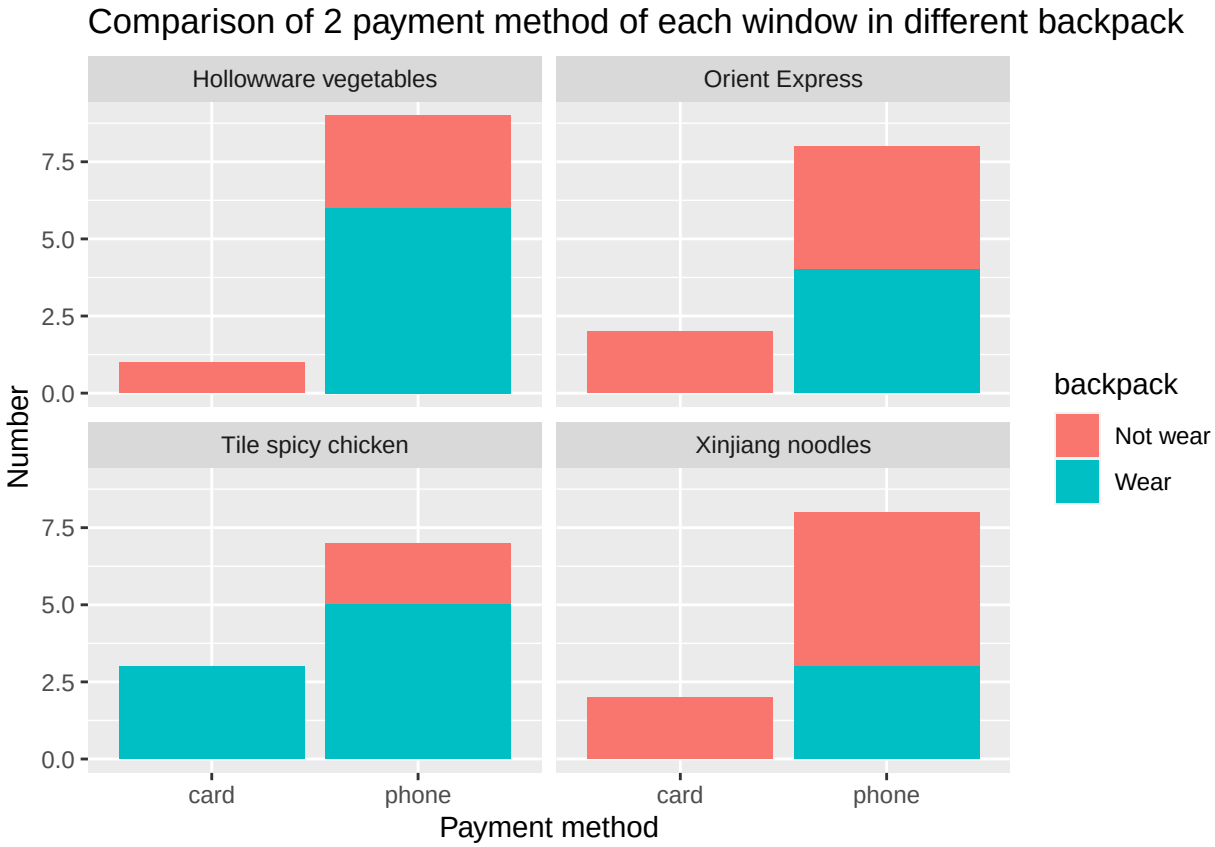
```
      title = "Comparison of 2 payment method in different backpack")
```




```
## Comparison of 4 windows using facet function
# different gender
ggplot(data = Data1) +
  geom_bar(mapping = aes(x = pay_way, fill = gender)) +
  facet_wrap(~ window, nrow = 2) +
  # scale_y_continous(limits = c(0, 10), breaks = seq(0, 10, 1)) +
  labs(x = "Payment method", y = "Number",
       title = "Comparison of 2 payment method of each window in different gender")
```




```
# different backpack condition
ggplot(data = Data1) +
  geom_bar(mapping = aes(x = pay_way, fill = backpack)) +
  facet_wrap(~ window, nrow = 2) +
  labs(x = "Payment method", y = "Number",
       title = "Comparison of 2 payment method of each window in different backpack")
```



Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

```
library(pwr)
```

```
## Warning: package 'pwr' was built under R version 4.0.3
```

```
# f2 test
pwr.f2.test(u=2,v=37,sig.level=0.05,power=0.80)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 37
##              f2 = 0.2614587
##       sig.level = 0.05
##           power = 0.8
```

```
# sample size
# According to the sample's result, Pr(card)=32/40=0.8.
# So I believe the true Pr(card) is around 0.8.
n <- 0.8*(1-0.8)/0.05^2
```

In this test, I use pwr.f2.test() function to do power analysis, because I want to fit it in GLM. System calculates f2 = 0.2614587. Small effect is around 0.02; middle effect is around 0.15; big effect is around 0.35. 0.26 is between 0.15 and 0.35. My sample size is 40 and suggested size is 64. I think it can answer the first question, which payment method students prefer. But as for prediction, it is not enough. Any power analysis or sample size calculations is conditional on an assumed effect size, and this is something that is the target of the study and is thus never known ahead of time.

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

```
library(rstanarm)
```

```
## Loading required package: Rcpp
## This is rstanarm version 2.21.1
## - See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!
## - Default priors may change, so it's safest to specify priors, even if equivalent to the defaults.
## - For execution on a local, multicore CPU with excess RAM we recommend calling
##   options(mc.cores = parallel::detectCores())
```

```
set.seed(100)
M1 <- stan_glm(pay_way ~ gender + backpack, data = Data2,
               family=binomial(link="logit"), refresh = 0)
summary(M1)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       binomial [logit]
##  formula:      pay_way ~ gender + backpack
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 40
##  predictors:   3
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept)  0.8    1.0 -0.4   0.7   2.0
## gender       0.4    1.0 -0.9   0.5   1.7
## backpack     0.8    0.8 -0.3   0.8   1.8
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.8    0.1  0.7   0.8   0.9
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  4060
```

```
## gender        0.0  1.0  3862
## backpack      0.0  1.0  3109
## mean_PPD      0.0  1.0  4121
## log-posterior 0.0  1.0  1784
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

I set link = "logit" because I want to use logistic regression in this case. Why logistic regression? Since I want to know which way students like.There are just 2 ways and students should choose 1 of them. Outcome is binary outcome.

Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

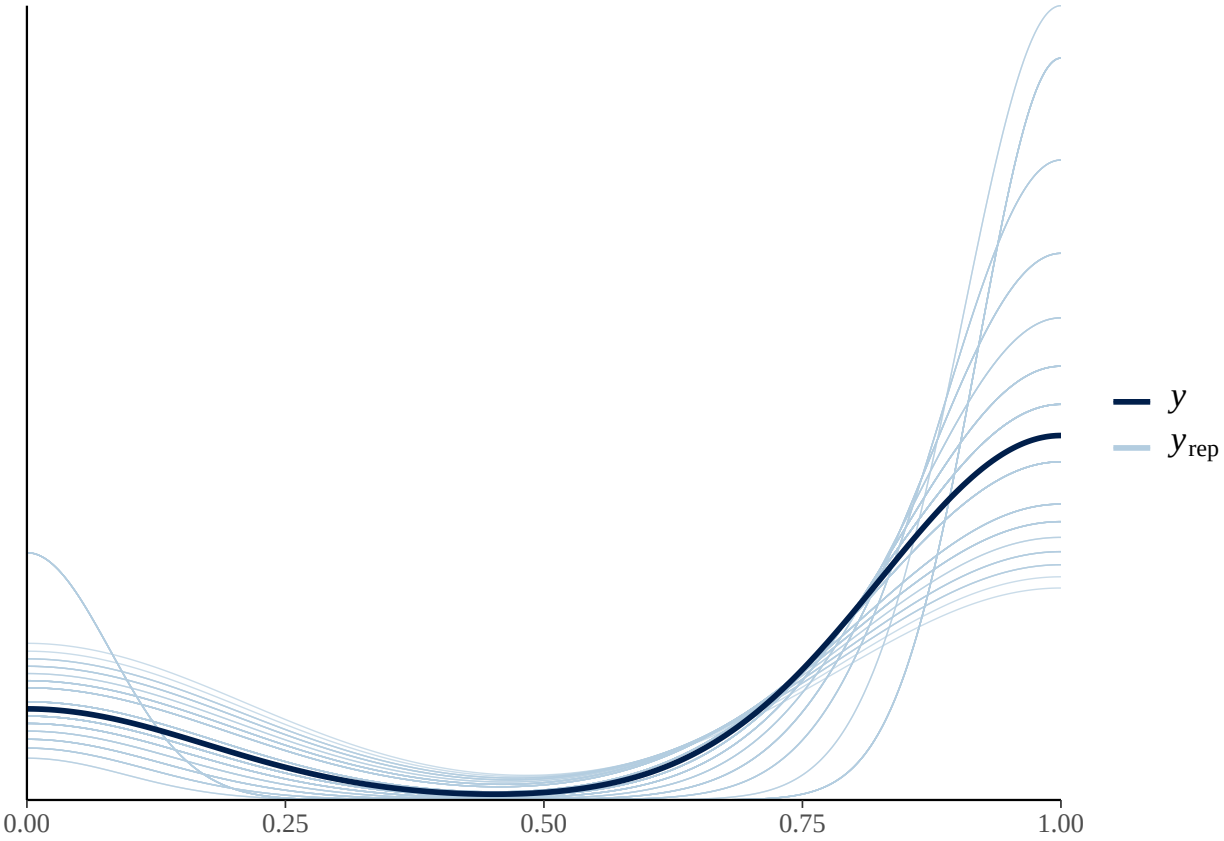
```
library(rstanarm)
library(bayesplot)
```

```
## This is bayesplot version 1.7.2
## - Online documentation and vignettes at mc-stan.org/bayesplot
## - bayesplot theme set to bayesplot::theme_default()
##    * Does _not_ affect other ggplot2 plots
##    * See ?bayesplot_theme_set for details on theme setting
```

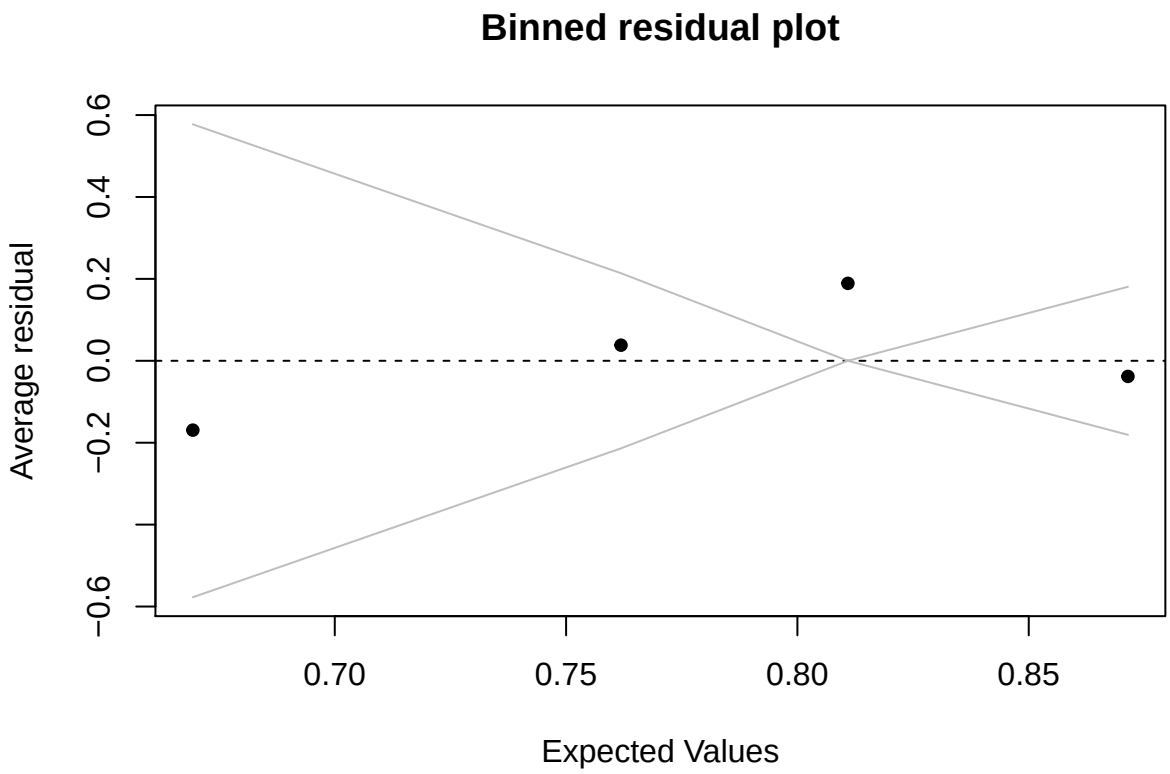
```
library(arm)
```

```
## Loading required package: MASS
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
## Loading required package: lme4
##
## arm (Version 1.11-2, built: 2020-7-27)
## Working directory is D:/MSSP/Rdata/678 Midterm exam
##
## Attaching package: 'arm'
## The following objects are masked from 'package:rstanarm':
##
##     invlogit, logit
```

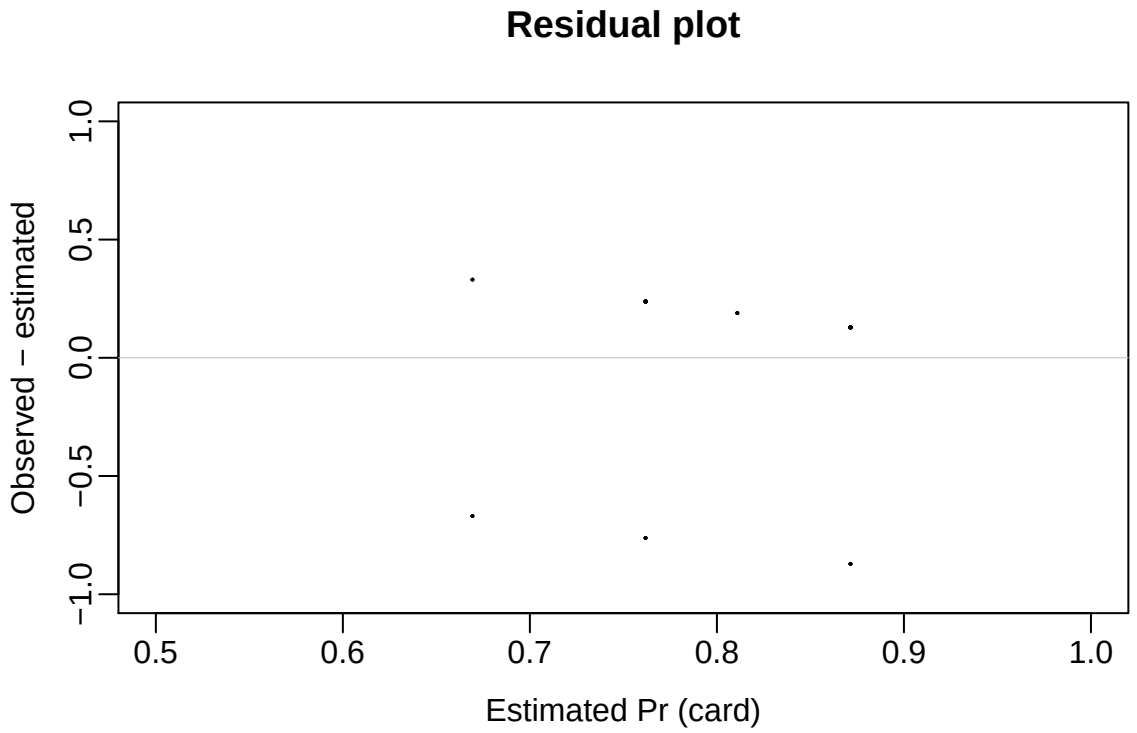
```r
# PPC distributions
post_M1 = posterior_predict(M1,draws=1000)
ppc_dens_overlay(post_M1[1:100,], y=Data2$pay_way)
```




```r
# binned residual plot
binnedplot(fitted(M1), resid(M1))
```




```
# residual plot
plot(c(0.5,1), c(-1,1), xlab="Estimated Pr (card)", ylab="Observed - estimated",
     type="n", main="Residual plot", mgp=c(2,.5,0))
abline(0,0, col="gray", lwd=.5)
points(fitted(M1), Data2$pay_way-fitted(M1), pch=20, cex=.2)
```




```
# error rate
error_rate <- mean((fitted(M1)>0.5 & Data2$pay_way==0) | (fitted(M1)<0.5 & Data2$pay_way==1))
error_rate
```

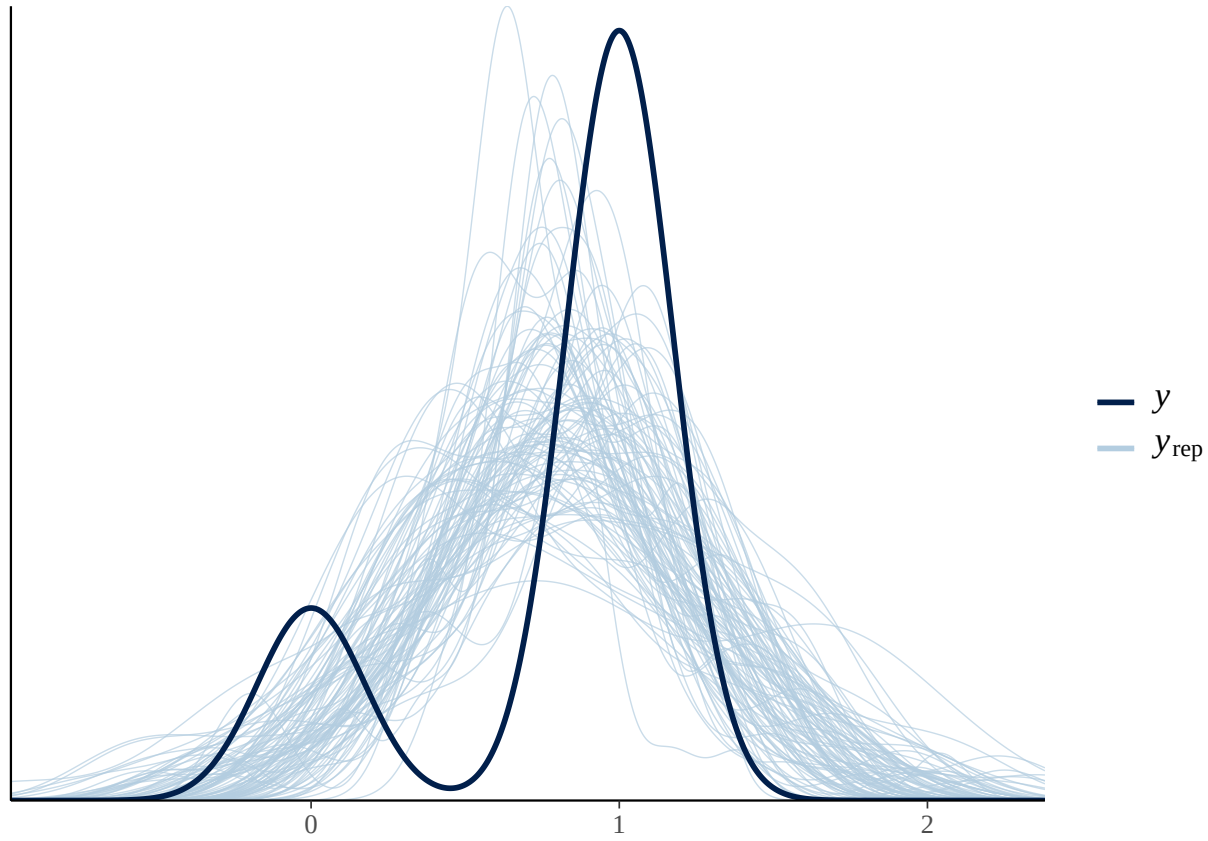
```
## [1] 0.2
```

```
# if use linear model
M2 <- stan_glm(pay_way ~ gender + backpack, data = Data2, refresh = 0)
print(M2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      pay_way ~ gender + backpack
##  observations: 40
##  predictors:   3
## ------
##             Median MAD_SD
## (Intercept) 0.7    0.2
## gender      0.1    0.2
## backpack    0.1    0.1
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.4    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
```

```
## * For info on the priors used see ?prior_summary.stanreg
```

```
post_M2 = posterior_predict(M2,draws=1000)
ppc_dens_overlay(post_M2[1:100,], y=Data2$pay_way)
```




```
# LOO
loo(M1)
```

```
##
## Computed from 4000 by 40 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo    -22.8 4.3
## p_loo         3.3 0.8
## looic        45.6 8.5
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     38    95.0%   2097
##  (0.5, 0.7]   (ok)        2     5.0%   1581
##    (0.7, 1]   (bad)       0     0.0%   <NA>
##    (1, Inf)   (very bad)  0     0.0%   <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

```r
loo(M2)
```

```
##
## Computed from 4000 by 40 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    -23.9  5.0
## p_loo         4.1  0.9
## looic        47.8 10.1
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

The LOO estimated log score (elpd_loo) for the second model(M2) is lower than first model(M1).

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

```r
# Pr(card)=invlogit(0.8 + 0.4*gender + 0.8*backpack)
# Predict Pr(card) of female without backpack
pre1 <- data.frame(gender=0, backpack=0)
epred1 <- invlogit(posterior_linpred(M1, newdata = pre1))
mean(epred1)
```

```
## [1] 0.652567
```

```r
# Predict Pr(card) of female with backpack
pre2 <- data.frame(gender=0, backpack=1)
epred2 <- invlogit(posterior_linpred(M1, newdata = pre2))
mean(epred2)
```

```
## [1] 0.7789743
```

```r
# Predict Pr(card) of male without backpack
pre3 <- data.frame(gender=1, backpack=0)
epred3 <- invlogit(posterior_linpred(M1, newdata = pre3))
mean(epred3)
```

```
## [1] 0.7519526
```

```r
# Predict Pr(card) of male with backpack
pre4 <- data.frame(gender=1, backpack=1)
epred4 <- invlogit(posterior_linpred(M1, newdata = pre4))
mean(epred4)
```

```
## [1] 0.8613429
```

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

From the result, most people prefer to use mobile phone as their payment method. Gender and backpack will affect Pr(card), but not too much. As for gender, when other condition does not change, Pr(card|male) is around 10% higher than Pr(card|female). The main reason, by my observation, is probably that female like to decorate their student ID card. I saw these cards, which female used, with different stickers and cutting

ferrules. Since the card is beautiful, use it. On the other hand, male's card is normal.
As for backpack, when other condition does not change, Pr(card|with backpack) is around 10% higher than Pr(card|without backpack). The first reason, which I mentioned at beginning, is if students just finished classes, they must carry their ID card. The second reason may be the only way of entering library and laboratory is to use ID card. These students may go to library after finishing lunch. So they also carried ID card.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study. Concerns: 1. The sample size is too small. I observed 40 students. But my university has more than 20000 students. 2. There is a big difference in proportion between female and male. Generally speaking, 50:50 is good. But my university is famous for the amazing male-female ratio, which is 7:1. So, it's hard to get good sample. 3. You may have noticed that I did not use multilevel regression. Although I divided my observation into 4 groups, which are 4 food windows, these 4 windows have no difference in payment method. The differences are price, popularity, and time of offering food. 4. I did not have deep understanding of the result of power analysis and validation part.

Future: 1. Using internet is an option. Website "CSDN" has many good understanding. 2. The most fast way to get a framework of new knowledge is to talk with my peers. I like this way very much. Since this exam is not allowed to talk to each other, when I met problems, Internet and PPT are my helpers. 3. Do project with classmates as much as I can. That really helps a lot.

Comments or questions

If you have any comments or questions, please write them here.