Midterm Exam

Donghao Xue

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.

```
coin = read.csv("C:/Users/Daniel Xue/Desktop/678/coin_data.csv", encoding = "UTF-8")
colnames(coin) = c("0.01_yuan","0.05_yuan","0.1_yuan","0.5_yuan")

coin
```

```
##   0.01_yuan 0.05_yuan 0.1_yuan 0.5_yuan
## 1      8.12     11.06     8.72     8.75
## 2      5.72      8.73     7.95     9.33
## 3      8.36      7.62     8.45     6.59
## 4      6.40     10.22     6.08     7.73
## 5      6.43      9.00     9.12     6.63
## 6      5.83      9.98     9.59     7.21
## 7      6.78      7.63    10.99     9.19
## 8      7.25      9.92     8.20     8.66
```

There are four types of coins which value 0.01, 0.05, 0.1 and 0.5 separately. For each coin, I recorded the time of spinning in seconds for eight times.

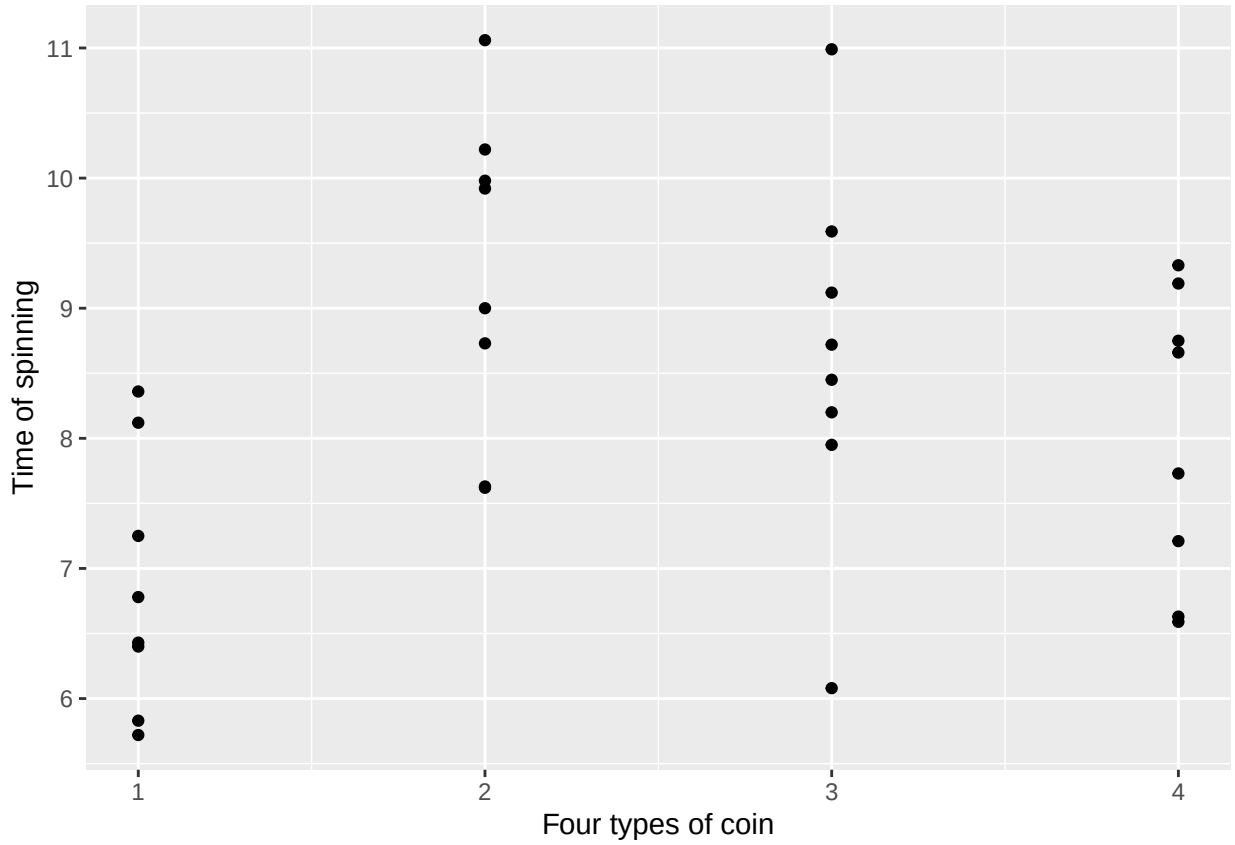
I want to figure out which type of coin can spin for a longer time.

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.

```
x = c(rep(1,8), rep(2,8), rep(3,8), rep(4,8))
y = c(coin$`0.01_yuan`, coin$`0.05_yuan`, coin$`0.1_yuan`, coin$`0.5_yuan`)

ggplot(data.frame(x,y), aes(x,y)) + geom_point() + xlab("Four types of coin") + ylab("Time of spinning")
```



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.

```
#There are four samples in this problem and I will compare the first two types of coin

effect_size = (mean(coin$`0.01_yuan`) - mean(coin$`0.05_yuan`))/sd(c(coin$`0.01_yuan`,coin$`0.05_yuan`))

pwr.t.test(n = 8, d = NULL, sig.level=0.05,power=0.8,type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 8
```

```
##               d = 1.50665
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
effect_size
```

```
## [1] -1.461035
```

Based on the calculation, the effect size is about 1.46 which is close to the result from power analysis. I think the sample size is enough for the problem. The published result will cause overestimation, so I should not use the effect size from the fitted model.

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.

```
f1 = stan_glm(y ~ factor(x)-1, data = data.frame(x,y), refresh=0)

print(f1, digits=2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      y ~ factor(x) - 1
##  observations: 32
##  predictors:   4
## ------
##            Median MAD_SD
## factor(x)1 6.86   0.42
## factor(x)2 9.23   0.44
## factor(x)3 8.63   0.43
## factor(x)4 7.99   0.44
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 1.22   0.16
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

I choose to use a linear regression model for this problem, since the data has only one predictor which is the type of coins and the outcome is not binary.

Validation (10pts)

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

```
cbind(mean(coin$`0.01_yuan`), mean(coin$`0.05_yuan`),mean(coin$`0.1_yuan`),mean(coin$`0.5_yuan`))
```

```
##         [,1] [,2]   [,3]    [,4]
## [1,] 6.86125 9.27 8.6375 8.01125
```

After calculating the mean spinning time for each coin, the result is similar to the coefficients of the model.

Inference (10pts)

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

```
p1 = predict(f1, newdata = data.frame(x = factor(1)))

p2 = predict(f1, newdata = data.frame(x = factor(2)))

p3 = predict(f1, newdata = data.frame(x = factor(3)))

p4 = predict(f1, newdata = data.frame(x = factor(4)))


cbind(mean(p1),mean(p2),mean(p3),mean(p4))
```

```
##          [,1]    [,2]     [,3]     [,4]
## [1,] 6.85064 9.23422 8.632753 7.991729
```

From the prediction we can see the mean of prediction value is similar to the original data.

Discussion (10pts)

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

Based on the result of the model, we can see the four cosfficients corresponding to the four types of coin, it implies that the second type of coin which values 0.05 can spin for a longer time, which has an average spinning time of 9.24s.

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.

In this analysis, I only use one predictor to build the model, so I can improve the accuracy of the result by adding more predictors such as the weight of the coin and the thickness of the coin. In the process of getting the data, I recorded the time by hand so it will cause measurement error. I may reduce it by increasing the times of measurements, and taking average.

Comments or questions

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