Midterm Exam

Ruxin Liu

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.

The photo data is about the number of pictures taken by cell phones among different groups of people and it was collected through surveys. The participants are split into one of the working group (working physically at the company), the remote working group (working remotely from distance), the learning group (learning physically at the campus) or the remote learning group (learning remotely from distance) based on their current status. The number of pictures is collected as a numeric variable. The gender is collected as a binary variable where F is female and M is male. The cell phone operating system indicates whether the participant uses ios or Android. The location variable tells the information about where the participants currently are. There are 5 observations per group and the data has 20 observations in total with no missing values. The question I am interested in comparing is that does the number of pictures taken by cell phones varies among different groups of people.

In order to avoid potential effects due to various ages, all chosen participants were born between the year 1996 and the year 1999, which can be considered as in the same age group. All the participants are selected from my contact list and are asked individually to count the number of pictures (including screenshots) they took with their cell phones for one particular week (from 20/10/18 to 10/10/25). All the participants can be considered as independent since none of them works for the same company or studies at the same school.

```
library(kableExtra)
# Load in the data, the data is uploaded in the gitHub Repo
photo <- read.csv("Data Ruxin Liu.csv")
```

```r
# Rename some of the columns
colnames(photo)[3] <- "Picture_number"
colnames(photo)[5] <- "Operating_system"
colnames(photo)[6] <- "Location"
kable(photo[c(1, 6, 11, 16), ])
```

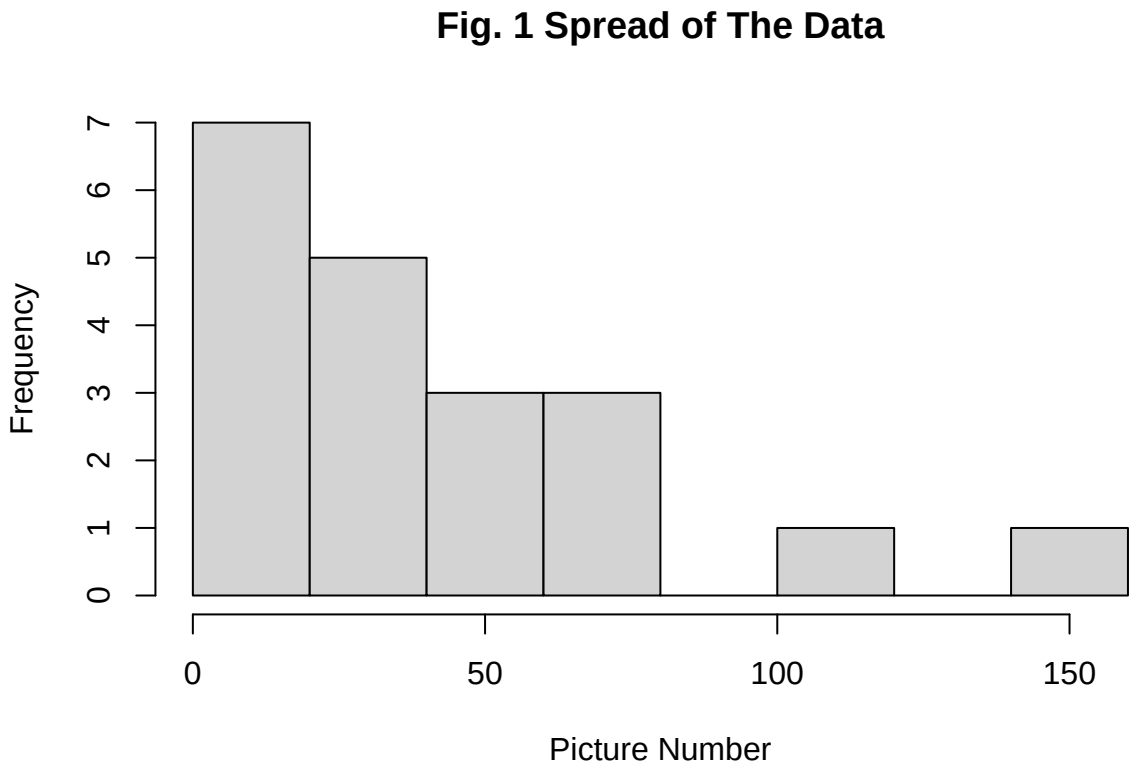
	Group	Participants	Picture_number	Gender	Operating_system	Location
1	Working	1	69	F	Android	China – Shanghai
6	Remote working	1	33	F	ios	China – Shanghai
11	Learning	1	4	M	ios	China – Shanghai
16	Remote learning	1	8	F	Android	China – Shanghai

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.

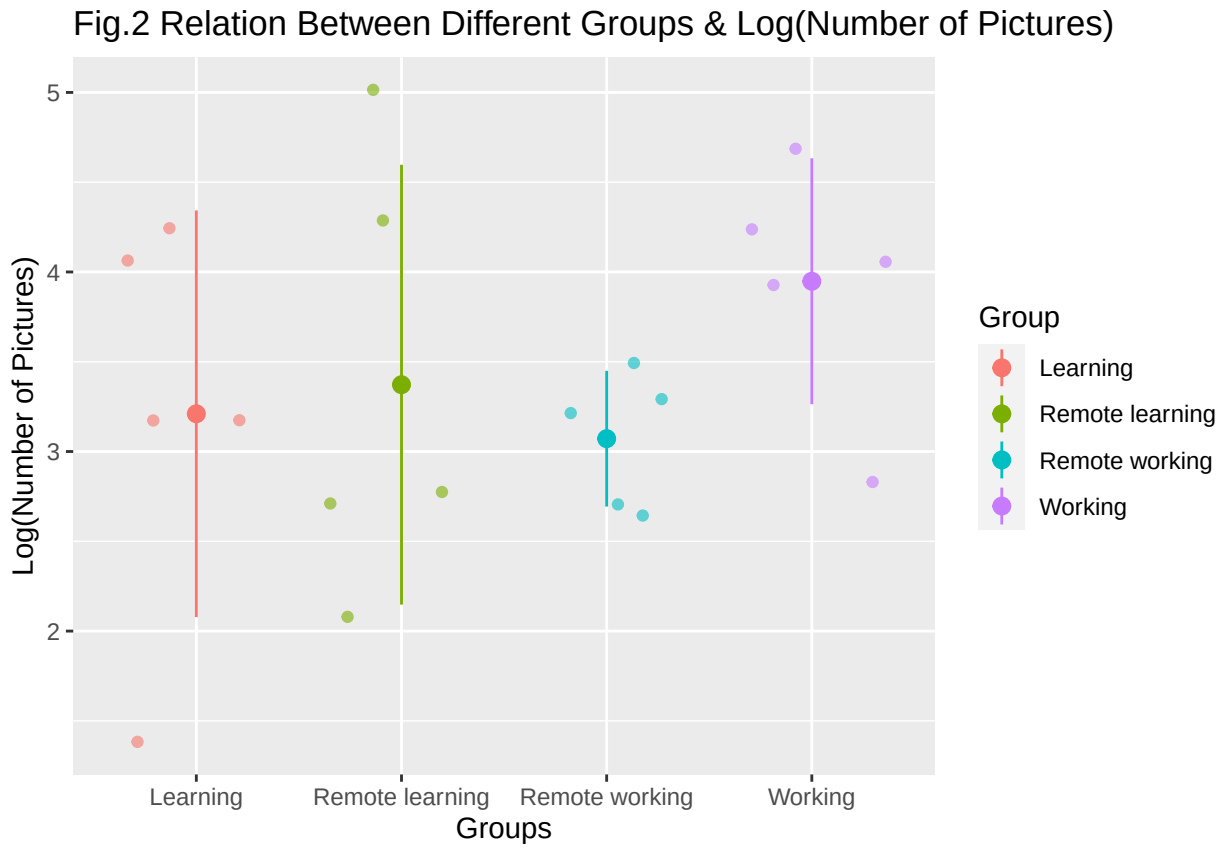
From Fig.1 below, it is very clear that the data is right-skewed, which suggests that logarithmic transformation is needed for the analysis. Also, since that the value of 0 (no photo is taken) is also meaningful, there is no need to center the data in this case.

```r
hist(photo$Picture_number, xlab = "Picture Number", main = "Fig. 1 Spread of The Data")
```



From Fig.2 below, the numbers of pictures in log scale for each group are plotted and the vertical line indicates the range of mean +/- 1 standard deviation. The learning group and the remote learning group don't show big differences, while the remote working group and the working group both seem to be distinct from the other two groups. This plot shows some ideas about my contrast of interest among different groups, which will be further analyzed.

```r
library(dplyr)
library(ggplot2)
# Create new variable after log transformation -- log(Picture_number)
photo$Log_number <- log(photo$Picture_number)
# Calculate the mean and standard deviation for each group
photo_mean <-photo %>%
  group_by(Group) %>%
  summarize(mean = mean(Log_number), sd = sd(Log_number))
ggplot(photo, aes(x = Group, y = Log_number, col = Group)) +
  # Avoid overplotting -- show all 20 points on the graph
  geom_jitter(alpha = 0.6) +
  ggtitle("Fig.2 Relation Between Different Groups & Log(Number of Pictures)") +
  labs(x = "Groups", y = "Log(Number of Pictures)") +
  geom_pointrange(data = photo_mean, aes(x = Group, y = mean,
                                         ymin = mean - sd, ymax = mean + sd))
```



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.

The result of the power analysis using 80% power, 95% significance level and sample size per group of 5 is showed below and the level of effect size is inferred to be 2.024, which is very large.

```r
# Perform power analysis -- infer the effect size
library(pwr)
# The data contains 4 different groups and each has a sample size of 5.
pwr.t.test(n = 5, d = NULL, sig.level = 0.05, power = 0.8, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 5
##               d = 2.024439
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The result of the power analysis using 80% power, 95% significance level and effect size of 0.5 is showed below to calculate the appropriate sample size in each group. Since there are no previous or relevant studies to refer, the effect size needs to be assumed properly. My initial guess is that the number of pictures taken by cell phone among different groups vary, but not differing by a lot. Therefore, I decide to set the effect size level as medium, which is 0.5 suggested by Cohen. Based on the result below, the sample size for each group should be 64, which is much higher than the sample size I currently have. Therefore, my sample size was not enough for the problem at hand.

```r
# Perform power analysis -- find the sample size
pwr.t.test(n = NULL, d = 0.5, sig.level = 0.05, power = 0.8, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The insufficient sample size suggests that this study is very likely to be an underpowered study, which is very harmful. If the effect size from the fitted model is used for power analysis in this case, the result could be overestimated and also very possible to lead to Type S error and Type M error. Therefore, for the purpose of accuracy, the effect size from the fitted model should not be used.

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.

The regression model that best fits this data is the linear regression model with transformation. For this study, the response (Picture_number) is a continuous variable, so my first thought is to fit a multilevel linear regression (lmer(Picture_number ~ factor(Group) + Gender + (1 | Operating_system), data = photo)). In the data, there are no repeated measurements for the individual, but it seems to have some nested structure for the gender or the operating system. However, I realized that there is no evidence that these two groups are natural hierarchy, unlike school level or country level, and also no group-level predictors are collected in the data. In addition, as Gelman and Hill stated multilevel models will reduce to clasical regressions when

the group only has two levels (Data Analysis Using Regression and Multilevel/Hierarchical Models: pg 275). Therefore, instead of fitting multilevel models with only 2 levels, I decide to fit a linear regression model with the binary variables Gender and Operating_system included as predictors.

As discovered from the EDA, log transformation needs to be performed to the response variable. Therefore, for the linear model I fitted, the response variable is Log_number and the predictors are Group, Gender and Operating_system. The reason why the Location variable is not included is that for 20 observations there are 10 different locations, therefore adding this information is not very representative due to the small sample size. The model output is shown below:

```
library(pander)
# length(unique(photo$Location))
fit_photo <- lm((Log_number) ~ factor(Group) + Gender + Operating_system, data = photo)
pander(summary(fit_photo))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.79	0.6792	4.108	0.001066
factor(Group)Remote learning	0.3045	0.6145	0.4956	0.6279
factor(Group)Remote working	-0.1314	0.5879	-0.2235	0.8264
factor(Group)Working	0.7453	0.5879	1.268	0.2256
GenderM	-0.6408	0.4938	-1.298	0.2154
Operating_systemios	0.6768	0.5599	1.209	0.2468

Table 2: Fitting linear model: (Log_number) ~ factor(Group) + Gender + Operating_system

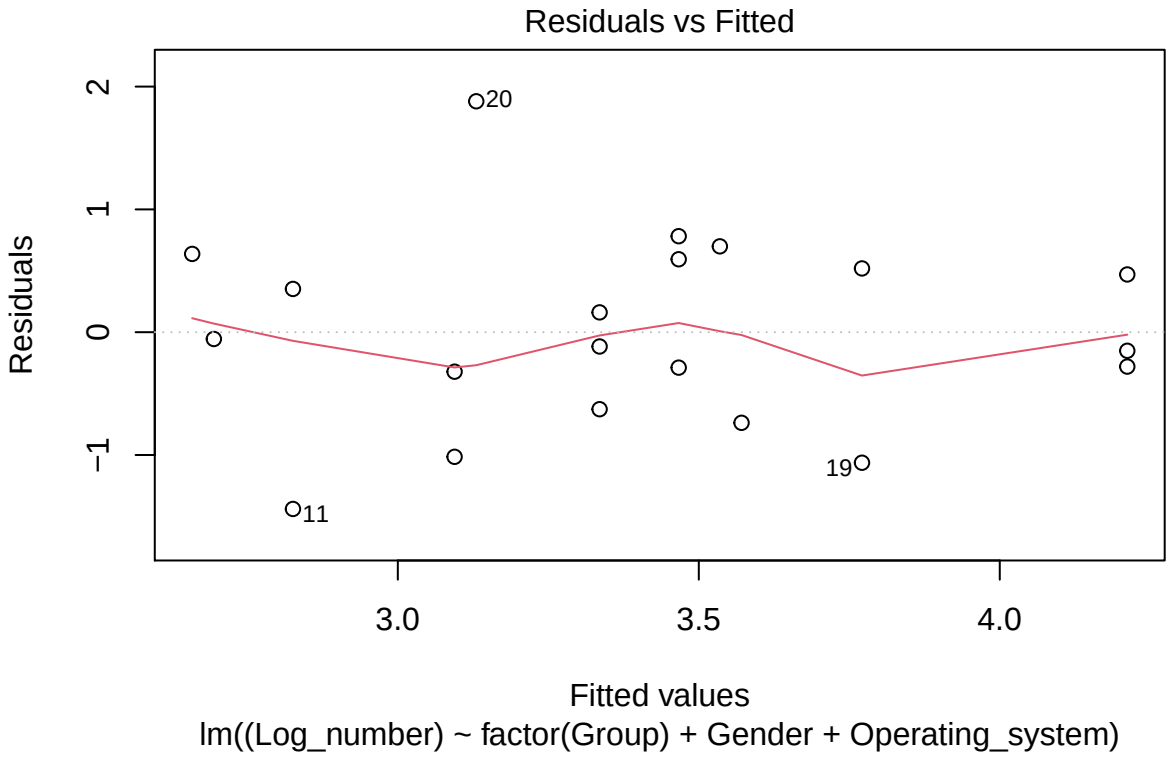
Observations	Residual Std. Error	R^2	Adjusted R^2
20	0.9067	0.2717	0.01155

Validation (10pts)

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

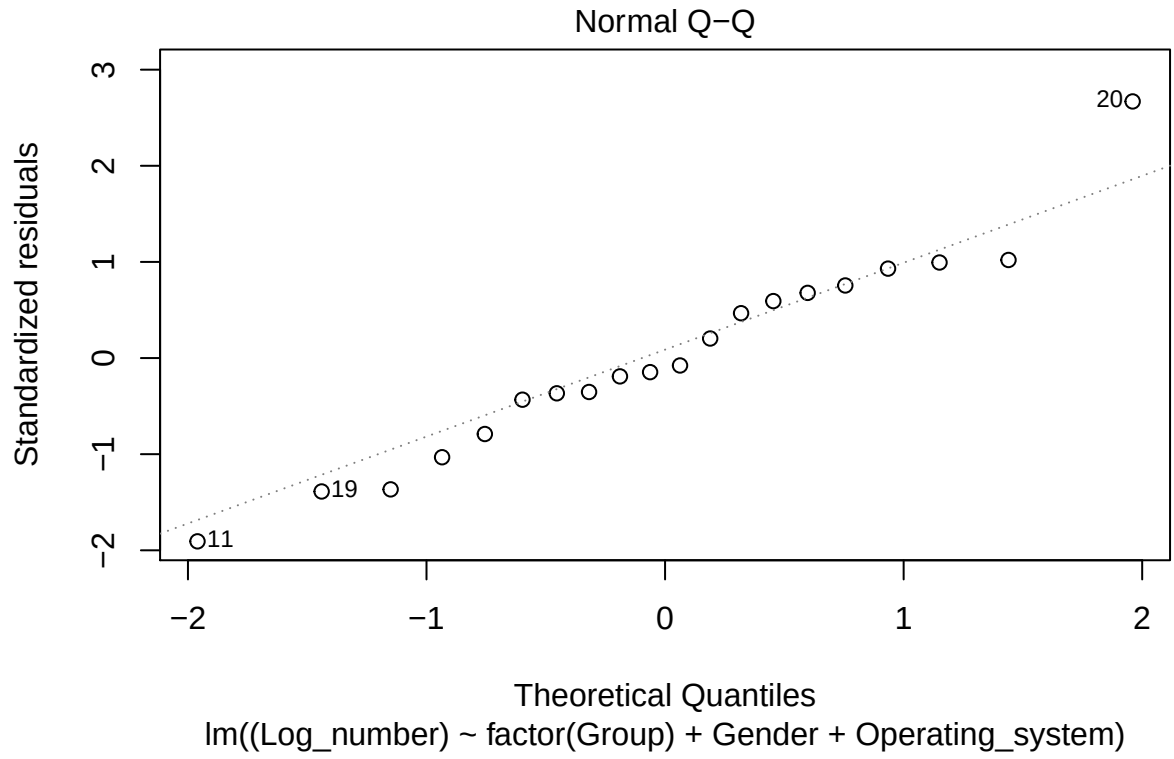
One assumption of linear regression is homoscedasticity, which can be tested by the Residuals vs Fitted Plot. From the plot below, although the points are not perfectly well-spread, there is no distinct pattern or cluster, this assumption is met.

```
plot(fit_photo, which = 1)
```



Another assumption of linear regression is normality, which can be tested by the Normal Q-Q Plot. From the plot below, it is clear that most of the residuals fall very close to the straight line, which suggests that this assumption is met.

```
plot(fit_photo, which = 2)
```



Also, I can check whether this model is appropriate by performing cross-validation. I fit another linear regression without the log transformation and without the Gender and Operating_system variables. The model output is shown below:

```
fit_photo_2 <- lm((Picture_number) ~ factor(Group), data = photo)
pander(summary(fit_photo_2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36	16.65	2.162	0.04611
factor(Group)Remote learning	16.4	23.55	0.6965	0.4961
factor(Group)Remote working	-13.2	23.55	-0.5606	0.5829
factor(Group)Working	24.6	23.55	1.045	0.3117

Table 4: Fitting linear model: (Picture_number) ~ factor(Group)

Observations	Residual Std. Error	R^2	Adjusted R^2
20	37.23	0.1616	0.004428

```
# Calculate the mean square error
 mean(fit_photo$residuals^2)
```

```
## [1] 0.5755138
```

```
 mean(fit_photo_2$residuals^2)
```

```
## [1] 1108.96
```

By comparing the two models, the first model fit_photo has a R-squared value of 0.2717, while the second model fit_photo_2 has a R-squared value of 0.1616. The larger value that fit_photo has suggests that the first model can explain more variations in the data, and the adjusted R-square shows the same result. Also, the first model has a much smaller mean square error comparing to the second model, which suggests that the fit_photo is a more appropriate model for the data.

In conclusion, fit_photo (lm((Log_number) ~ factor(Group) + Gender + Operating_system, data = photo)) is an appropriate model for the data, because it met the important assumptions for linear regression and also performed better in the cross-validation.

Inference (10pts)

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

The linear regression model fitted for the data is: log(y) = 2.79 + 0.30 remote_learning - 0.13 remote_working + 0.75 working - 0.64 male + 0.68 ios. The intercept is 2.79, which means that for a female uses Android system and from the learning group (reference group), the estimated picture she will take for a week is 16 ($e^{2.79}$). When keeping all other variables constant, the remote learning group is estimated to take 35% more pictures compared to the learning group ($(e^{0.3} - 1) * 100$). When keeping all other variables constant, the remote working group is estimated to take 12% less pictures compared to the learning group ($(e^{-0.13} - 1) * 100$). When keeping all other variables constant, the working group is estimated to take 112% more pictures compared to the learning group ($(e^{0.75} - 1) * 100$).

Although from the interpretation, it seems that there are clear difference between the number of pictures taken by different groups, except for the intercept all other predictors are not statistically significant on the

95% level. The results of the 95% confidence interval for the coefficients are shown below and all coefficients for the group indicator accross 0.

```
kable(confint(fit_photo))
```

	2.5 %	97.5 %
(Intercept)	1.3331185	4.2463898
factor(Group)Remote learning	-1.0134796	1.6225579
factor(Group)Remote working	-1.3922480	1.1294483
factor(Group)Working	-0.5155696	2.0061266
GenderM	-1.6999810	0.4183117
Operating_systemios	-0.5241113	1.8778068

Discussion (10pts)

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

The model fit for the photo data is the linear regression, where the response is the log-transformed Picture_number and the predictors include Group (factor with 4 levels), Gender and Operating_system. From the model output, none of the coefficients of the group indicators are statistically significant on a 95% confidence level. Although p-value can be misleading sometime, since the assumptions of the linear model are valid in this case, p values could be used as reference. The very large p-value of the coefficients suggest that there is no difference among the groups. However, since the sample size is very small and the sample may not be representative, the study possibly have some errors, which make the results less convincing.

In conclusion, the answer to the question of interest of this study is that the number of pictures taken by cell phone for a week does not vary among groups. However, in order to make the results more accurate and persuasive, better experiment design and better sample collection are needed to increase the power of the study.

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.

1. Random Sampling

This photo data is collected by surveys and the participants are chosen from my contact list. This cannot be considered as random sampling, because I will tend to pick friends that are closer to me and also there is clear gender difference among my participants. I intentionally select people from the same age group, however, I realized that age can also be an important variable to include in the analysis. This concern could be fixed by using public survey website online to collect data or by interviewing random people from public regardless of their gender and age.

2. Sample Size

As the power analysis shows, the sample size in each group is way too small and the study is very likely to be a harmful underpowered study. In future studies, sample size should be increased by collecting data from more randomly selected participants, and the number of participants in each group does not need to be the same.

3. Measurement Time Range

In this data, each participant was asked about the number of pictures they took with cell phones in a whole week. However, reducing the time range to 5 business days may be a better choice due to the grouping. For example, someone who is working and someone who is studying show different status from Monday to Friday, but may not be that different during weekends. This concern can be improved by reformatting the survey question and can even ask participants to only count pictures they took when they were working or studying, which may be very challenging to do.

4. Multilevel Regression Model

I believe that a multilevel model will be better to answer my question of interest than the simple linear model . Unfortunately, some of the key information are missing from the collected data. For my question, it is possible to find a natural cluster, such as the company or the school, but is relatively hard to perform. Instead, repeated measurements could be a more applicable solution. In future studies, I could ask the participants to provide the number of pictures they took for some consecutive weeks(from Monday to Friday), which allows me to fit the multilevel model.

Comments or questions

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