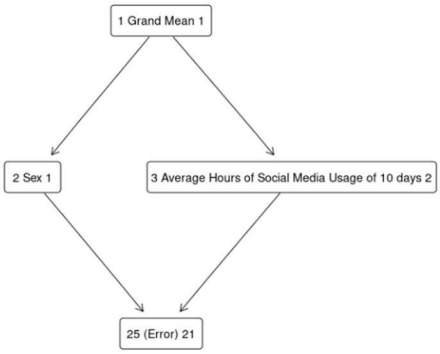
Statistical Investigation & Data Analysis Project

Sejun Song

✓ Set-up R

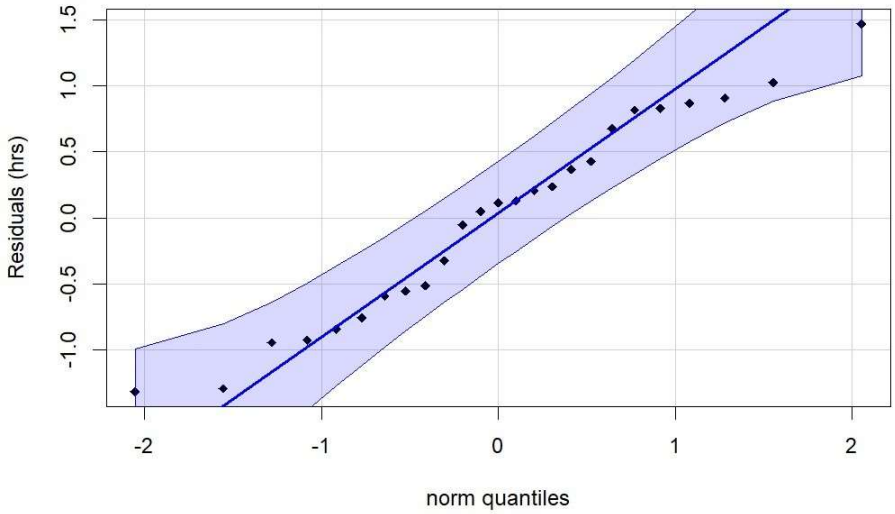
```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
options(knitr.kable.NA = "")
options(contrasts = c("contr.sum", "contr.poly"))
packages <- c("tidyverse", "knitr", "kableExtra",
"parameters", "hasseDiagram", "DescTools")
library(zoom)
library(ggplot2)
library(tidyverse)
library(tidyr)
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")
```
```

✓ Load Data

```
```{r}
socialmedia <- read.csv(
  file = "C:\\Users\\yss06\\Downloads\\Social Media Data.csv",
  header = TRUE,
  sep = ","
)

socialmedia$average_usage_hrs_10days <- factor(
x = socialmedia$average_usage_hrs_10days,
levels = c("High","Moderate", "Low")
)
## Set the block as a "factor"
socialmedia$Sex <- as.factor(socialmedia$Sex)


```
```

✓ Introduction and Background

- Social media became a huge culture nowadays. Everybody has their own smartphone and people do various things through smartphones such as social media, texting, calling, watch videos. Social Media became a hub of information, entertainment, and sharing their life with others. However, a lot of college students say they lack sleep. As college students, we spent a lot of time watching YouTube videos to learn things they could not get in class and to entertain ourselves and use Instagram and Facebook to socialize with people around the world so that we wanted to see how time spent on the three most popular social media (Instagram, Facebook, YouTube) affect to time of sleep for Penn State University students. Furthermore, we wanted to see how time spent on social media and time spent on sleeping can relate differently to sex and we will explain our study design and why we used it in the next section.

✓ Study Design and Methodology

- To investigate our research question, we have designed an experimental study. We asked for ten days of battery usage data for the three most popular social media from their smartphone and the time they slept from 40 Penn State students. We recorded their sex (M, F), mean time of ten days they spent on the three most popular social media, and mean time of ten days they spent sleeping. There are three levels of social media usage (Low, Moderate, High). Students who spent social media less than 5 hours are considered low, between 5 hours and 6 hours be moderate, and more than 6 hours are High. To answer our research question, we decided to use ANOVA with a block as the most appropriate method. We used sex as a block in our ANOVA with a block method. We used blocking by sex in ANOVA to pair and see how time spent on social media and time spent on sleeping can relate differently to sex.

✓ EDA

```
```{r Fitting Model}
sns_model <- aov(
formula = average_hrs_sleep_10days ~ average_usage_hrs_10days + Sex,
data = socialmedia
)


```
```



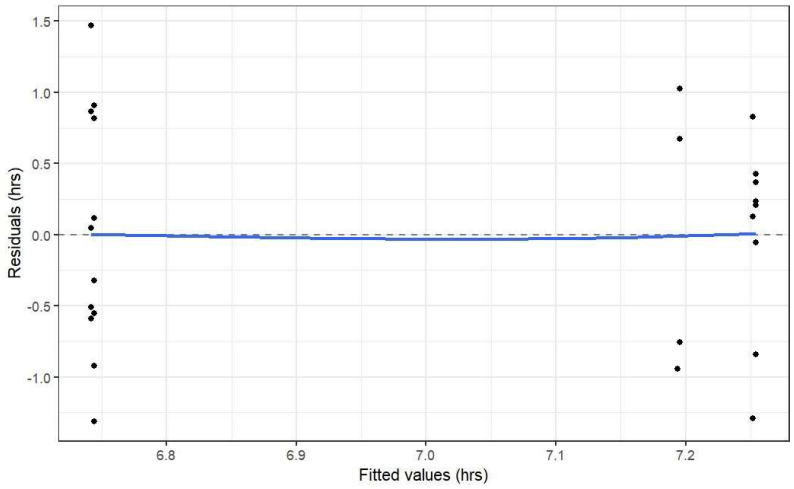
✓ Assumption

```r
car::qqPlot(
x = residuals(sns_model),
distribution = "norm",
envelope = 0.90,
id = FALSE,
pch = 18,
ylab = "Residuals (hrs)"
)
```



- Figure above shows the QQ plot for our residuals with a 90% confidence envelope. I don't see any obvious deviation of the points fall outside of the envelop, but there are two potential outliers on both side of the envelope, one on the first strip and the fifth strip from the left. However, the points most likely fall on a 45-degree reference line, showing a fair amount of variation. I can tell the variation of points on the blue line between the second and the fourth strip from the left form a satisfying, balanced pattern.

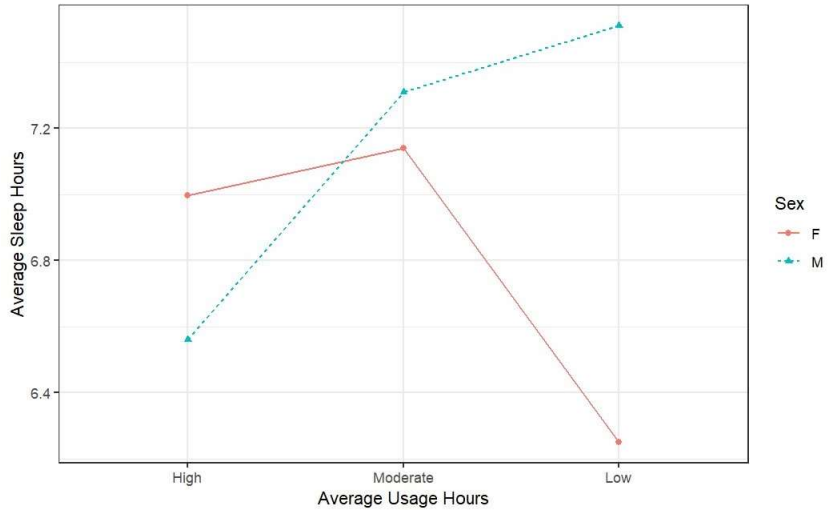
```r
# Homoscedasticity
ggplot(
data = data.frame(
residuals = residuals(sns_model),
fitted = fitted.values(sns_model)
),
mapping = aes(x = fitted, y = residuals)
) +
geom_point(size = 1.5) +
geom_hline( ## Adds reference line at zero
yintercept = 0,
linetype = "dashed",
color = "grey50"
) +
geom_smooth( ## Adds the smoothed line
formula = y ~ x,
method = stats::loess,
method.args = list(degree = 1),
se = FALSE,
size = 1
) +
theme_bw() +
xlab("Fitted values (hrs)") +
ylab("Residuals (hrs)")
```



- Based on the Tukey-Anscombe plot for the assumption of Homoscedasticity, I notice that the fifth strip from the left shows the least amount of variation while the first and little right of fifth strip shows the most. I can see an unbalanced, discernible patterns of variation of points on the fifth strip on the fitted values of 7.2 while I don't see any unbalanced patterns of variation for the first and last strips from the left. For the blue line across the variations, I find proper patterns, and the blue line is horizontal, showing the satisfaction of Homoscedasticity.
- For the issue of independence of observations, we know that 40 students and their sexuality are not randomly selected. but their social media usage is random, we can say that the independence of observation assumption is questionable.

✓ Assessing Interaction

```r
ggplot2::ggplot(
data = socialmedia,
mapping = aes(
x = average_usage_hrs_10days,
y = average_hrs_sleep_10days,
color = Sex,
shape = Sex,
linetype = Sex,
group = Sex
)
) +
stat_summary(fun = "mean", geom = "point") +
stat_summary(fun = "mean", geom = "line") +
ggplot2::theme_bw() +
xlab("Average Usage Hours") +
ylab("Average Sleep Hours") +
labs(color = "Sex") +
theme(
legend.position = "right"
)
```



- In the interaction plot, we don't see essentially the same behavior of average hours of social media usage in each sex; rather they across each other. Also, the lines are not even parallel. This indicates that there is any type of interaction between sex and average social media usage. Overall, the assumption of interaction between the subject (the sex) and the factor (the average hours of social media usage) is not being satisfied.

✓ ANOVA Table

```r
```{r}
parameters::model_parameters(
model = sns_model,
omega_squared = "partial",
eta_squared = "partial",
epsilon_squared = "partial"
) %>%
knitr::kable(
digits = 4,
col.names = c("Source", "SS", "df", "MS", "F", "p-value",
"Partial Omega Sq.", "Partial Eta Sq.", "Partial Epsilon Sq."),
caption = "ANOVA Table for Social Media Usage Study",
align = c('l',rep('c',8)),
booktab = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"),
font_size = 10,
latex_options = c("scale_down", "HOLD_position")
)
```
```

ANOVA Table for Social Media Usage Study

Source	SS	df	MS	F	p-value	Partial Omega Sq.	Partial Eta Sq.	Partial Epsilon Sq.
average_usage_hrs_10days	1.5202	2	0.7601	1.1023	0.3506	0.0081	0.095	0.0088
Sex	0.0000	1	0.0000	0.0000	0.9953	-0.0417	0.000	-0.0476
Residuals	14.4813	21	0.6896					

- For the interpretation of ANOVA Table, we can say that the average hours of social media usage accounted for about 1.1 times as much variation as residuals. This translates to average social media usage explaining around 0.88% of the total variation in students' sleep hour. Since our p-value is larger than our unusualness threshold (0.3506 < 0.05), we fail to reject the null hypothesis and decide to act as if social media usage does not impact the students' sleep hours. In particular, the average hours of social media usage accounts for around 0.81% of the variation in the students' sleep hours.

✓ Relative Efficiency

```{r Relative Efficiency}
block.RelEff(
aov.obj = sns_model,
blockName = "Sex",
trtName = "average_usage_hrs_10days"
)
```

```
[1] "The relative efficiency of the block, Sex, is 0.72."
```

- For the relative efficiency of the block, we can see that there is a statistically significant difference in the sleep hours that the sex gave due to the social media usage (F (2,21) = 1.1023, p value = 0.3506). The average social media usage accounted for just over 1.1 times as much variation even after accounting for sex effects. Under the null hypothesis of no effect due to social media usage, we would only anticipate seeing such an extreme F ratio, less than 1/100th of a percent of the time. Further, we can see from the rather large effect sizes, that amount of dosage accounts for around 0.88% of the variation in the sex' sleep hours on average of 10 days. The relative efficiency is approximately 0.72; thus, we would need about 1 time as many sexes for each social media usage as what we currently have to get the same level of information.

✓ Point Estimates

```{r}
pEst <- dummy.coef(sns_model)
pEst <- unlist(pEst[which(names(pEst) != "Sex")])
names(pEst) <- c(
"Grand Mean",
levels(socialmedia$average_usage_hrs_10days) # Using levels here will help stop the accidental
# mislabeling of estimates
)
data.frame("Estimate" = pEst) %>%
knitr::kable(
digits = 3,
caption = "Point Estimates from the Social Media Study",
booktabs = TRUE,
align = "c"
) %>%
kableExtra::kable_styling(
font_size = 12,
latex_options = c("HOLD_position")
)
```

Point Estimates from the Social Media Study

	Estimate
Grand Mean	7.064
High	-0.320
Moderate	0.189
Low	0.131

- For the point Estimates from the social media study, we would interpret the value of Grand Mean as 7.064 points per student; our entire sample accumulated 7.064 times as many points as sampled students.We can also see the factor level effects estimates. For High level of social media usage, they accumulated an additional -0.320 points per student whereas the Moderate level of social media usage only accumulated 0.189 points per student and the low level of social media usage accumulated 0.131 points per student. This shows that High level of social media usage indicates worse that baseline (GSAM).

✓ Effect Sizes

```r
```{r Effect Sizes}
anova.PostHoc(
aov.obj = sns_model, # Our aov output
response = "average_hrs_sleep_10days", # Our response variable
mainEffect = "average_usage_hrs_10days" # Our factor variable
) %>%
knitr::kable(
digits = 3,
caption = "Post Hoc Comparison Effect Sizes",
col.names = c("Pairwise Comparison","Cohen's d", "Hedge's g",
"Prob. of Superiority"),
align = 'lccc',
booktabs = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("condensed", "boardered"),
font_size = 12,
latex_options = "HOLD_position"
)

```
```

Post Hoc Comparison Effect Sizes

Pairwise Comparison	Cohen's d	Hedge's g	Prob. of Superiority
High vs. Moderate	-0.655	-0.629	0.322
High vs. Low	-0.510	-0.482	0.359
Moderate vs. Low	0.076	0.071	0.521

- Cohen's d and Hedge's g are both measures of the distance between the Sample Arithmetic Mean values for the two groups scaled by pooled standard deviations. The Probability of Superiority measure the percent of the time a randomly selected member of the first group will have the higher numeric value of the response than a randomly selected member of the second group.

- From the effect size table, we would say that there are rather small effects as Cohen's d and Hedge's g are all quite small. Further, the probability of superiority for each pairing is far from 0.5 (no practical effect). Just as effect sizes temper statistical significance, statistical significance moderates effect sizes. In these three cases, while there appears to be a small effect, there is not enough variation in those groups.

- We can say that there are -0.6555 standard deviations between the performance for using the high level of social media usage and the moderate level of social media usage. We also find that there are -0.510 standard deviations between using high level of social media usage and the low level of social media usage, and lastly, there are 0.076 standard deviations between the moderate level of social media usage and the low level of social media usage.

- For the Probability of Superiority, we can say only 32.2 % of the time we select a student given the high level of social media usage will that student have more average sleep hours than a selected student given the moderate level of social media usage. We also can see that 35.9% of the time we select a student given the high level of social media usage will affect a student have more average sleep hours than a selected student given the low level of social media usage. Additionally, 52.1% of the time we select a student given the moderate level of social media usage tend to have more average hours of sleep than a student given the low level of social media usage.

✓ Code Appendix

```
{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
options(knitr.kable.NA = "")
options(contrasts = c("contr.sum", "contr.poly"))
packages <- c("tidyverse", "knitr", "kableExtra",
"parameters", "hasseDiagram", "DescTools")
library(zoom)
library(ggplot2)
library(tidyverse)
library(tidyr)
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")

socialmedia <- read.csv(
  file = "C:\\Users\\yss06\\Downloads\\Social-Media-data.csv",
  header = TRUE,
  sep = ","
)

socialmedia$average_usage_hrs_10days <- factor(
x = socialmedia$average_usage_hrs_10days,
levels = c("High","Moderate", "Low")
)
## Set the block as a "factor"
socialmedia$Sex <- as.factor(socialmedia$Sex)

{r Fitting Model}
sns_model <- aov(
formula = average_hrs_sleep_10days ~ average_usage_hrs_10days + Sex,
data = socialmedia
)




{r Assumptions}

car::qqPlot(
x = residuals(sns_model),
distribution = "norm",
envelope = 0.90,
```

```
id = FALSE,
pch = 18,
ylab = "Residuals (hrs)"
)

# Homoscedasticity
ggplot(
data = data.frame(
residuals = residuals(sns_model),
fitted = fitted.values(sns_model)
),
mapping = aes(x = fitted, y = residuals)
) +
geom_point(size = 1.5) +
geom_hline( ## Adds reference line at zero
yintercept = 0,
linetype = "dashed",
color = "grey50"
) +
geom_smooth( ## Adds the smoothed line
formula = y ~ x,
method = stats::loess,
method.args = list(degree = 1),
se = FALSE,
size = 1
) +
theme_bw() +
xlab("Fitted values (hrs)") +
ylab("Residuals (hrs)")

ggplot2::ggplot(
data = socialmedia,
mapping = aes(
x = average_usage_hrs_10days,
y = average_hrs_sleep_10days,
color = Sex,
shape = Sex,
linetype = Sex,
group = Sex
)
) +
stat_summary(fun = "mean", geom = "point") +
stat_summary(fun = "mean", geom = "line") +
ggplot2::theme_bw() +
xlab("Average Usage Hours") +
ylab("Average Sleep Hours") +
```

```
labs(color = "Sex") +
theme(
legend.position = "right"
)

model_parameters_df  <- as.data.frame(parameters::model_parameters(
model = sns_model,
omega_squared = "partial",
eta_squared = "partial",
epsilon_squared = "partial"
))
knitr::kable(
model_parameters_df,
digits = 4,
col.names = c("Source", "SS", "df", "MS", "F", "p-value",
"Partial Omega Sq.", "Partial Eta Sq.", "Partial Epsilon Sq."),
caption = "ANOVA Table for Social Media Usage Study",
align = c('l',rep('c',8)),
booktab = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"),
font_size = 10,
latex_options = c("scale_down", "HOLD_position")
)

{r Relative Efficiency}
block.RelEff(
aov.obj = sns_model,
blockName = "Sex",
trtName = "average_usage_hrs_10days"
)

pEst <- dummy.coef(sns_model)
pEst <- unlist(pEst[which(names(pEst) != "Sex")])
names(pEst) <- c(
"Grand Mean",
levels(socialmedia$average_usage_hrs_10days) # Using levels here will help stop the accidental
# mislabeling of estimates
)
data.frame("Estimate" = pEst) %>%
knitr::kable(
digits = 3,
caption = "Point Estimates from the Social Media Study",
booktabs = TRUE,
align = "c"
```

```
) %>%
kableExtra::kable_styling(
font_size = 12,
latex_options = c("HOLD_position")
)


{r Effect Sizes}
anova.PostHoc(
aov.obj = sns_model, # Our aov output
response = "average_hrs_sleep_10days", # Our response variable
mainEffect = "average_usage_hrs_10days" # Our factor variable
) %>%
knitr::kable(
digits = 3,
caption = "Post Hoc Comparison Effect Sizes",
col.names = c("Pairwise Comparison","Cohen's d", "Hedge's g",
"Prob. of Superiority"),
align = 'lccc',
booktabs = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("condensed", "boardered"),
font_size = 12,
latex_options = "HOLD_position")
)
```