STA288 Final Project

Belal Hajjaj, Aaron Jin, Justin Yu, Brandon Luong, Modhar Al Qasser

2024-03-27

Author Contributions

BH, AJ, JY, BL, and MA - contributed to the development of the methodology
BH, AJ, JY, BL, and MA - contributed to the data collection
JY - contributed to R-markdown files JY - contributed to statistical analysis of the data
JY - contributed to drafting and writing the report

Introduction

Materials and Methods

Participants were sampled from the village of Arcadia on the island of Providence. Arcadia was chosen for its large population size of 4339, which encompasses a larger range of individuals than other villages, improving generalizability. Since there was no feasible method of obtaining a list of all individuals to take a simple random sample, a multistage sampling strategy was employed. Only adults at least 19 years of age were included to ensure participants were old enough to drink and complete study tasks.

Excel functions were used to generate a list of 100 house numbers out of the 1571 houses in Arcadia. From each selected house, the adults were numbered in order of appearance. R `runif()` was then used to generate a random number, and the corresponding resident was selected.

Participants who did not provide consent, empty houses, and duplicates were removed from the data, resulting in a final sample size of n = 90.

ANALYSIS (TEMPORARY)

Import data

```
raw <- read_csv("data.csv")
```

```
## Rows: 100 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Name, Alcohol_consumption, Forgetful
## dbl (3): House, Cards, Vocab
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Standardize data formatting

```
data <- raw %>%
  # keep only entries with data
  filter(Alcohol_consumption != "DUPLICATE" & Alcohol_consumption != "EMPTY" & Alcohol_consumption != "NO CONS
  # clean up formatting
  mutate(Name = str_to_title(Name), # change all names to title case

         # standardize alcohol consumption format
         Alcohol_consumption = case_when(
           str_detect(Alcohol_consumption, fixed("several times each season", ignore_case = T)) ~ "Several time
           str_detect(Alcohol_consumption, fixed("once or twice each season", ignore_case = T)) ~ "Once or twic
           str_detect(Alcohol_consumption, fixed("once or twice a year", ignore_case = T)) ~ "Once or twice a y
```

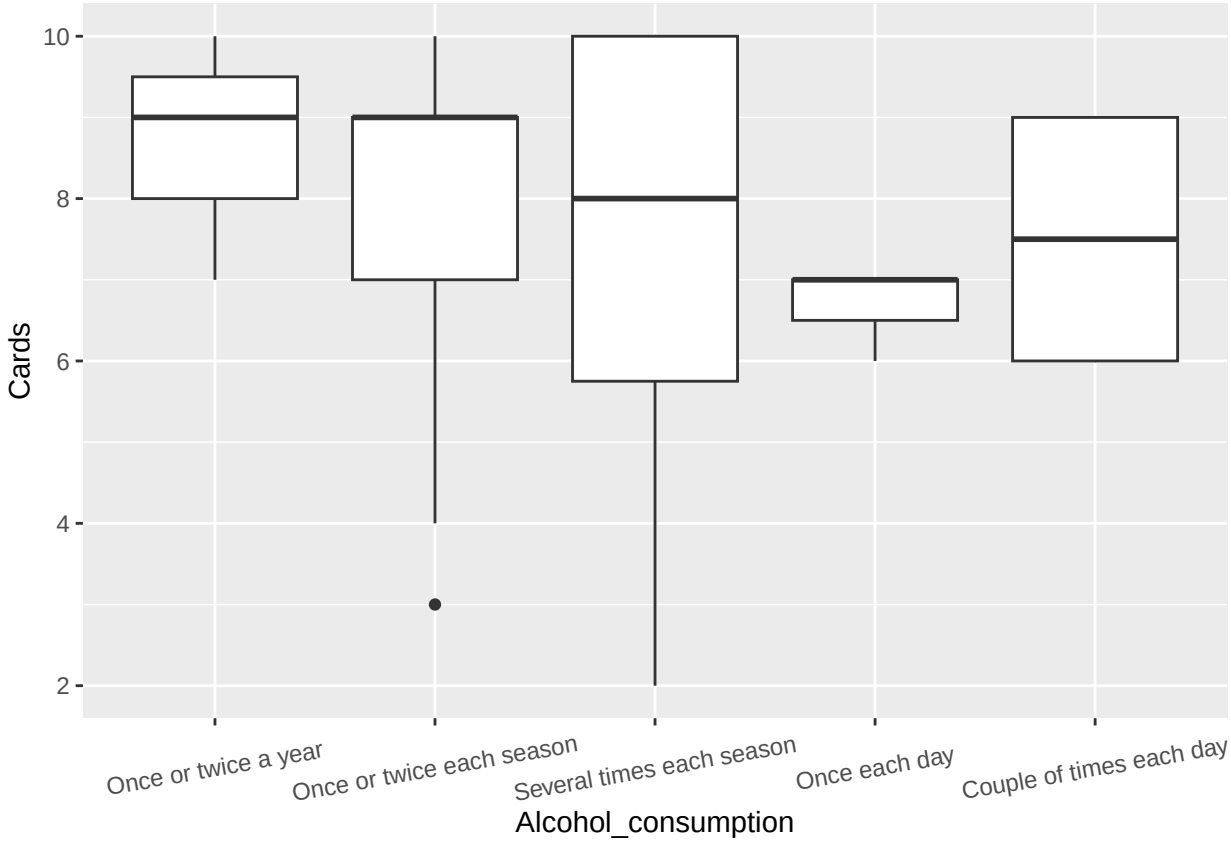
```
      str_detect(Alcohol_consumption, fixed("couple of times each day", ignore_case = T)) ~ "Couple of tim
      str_detect(Alcohol_consumption, fixed("drink each day", ignore_case = T)) ~ "Once each day"
    ),

    # standardize forgetfulness format
    Forgetful = case_when(
      str_detect(Forgetful, fixed("not at all", ignore_case = T)) ~ "Not at all",
      str_detect(Forgetful, fixed("a little", ignore_case = T)) ~ "A little",
      str_detect(Forgetful, fixed("moderately", ignore_case = T)) ~ "Moderately",
      str_detect(Forgetful, fixed("moderatly", ignore_case = T)) ~ "Moderately" # typo in one entry
    )
)
```

Preliminary Analyses *TEMPORARY: These results will be summarized using side-by-side boxplots, numerical summaries (mean, median, standard deviation, range), histograms, and stacked bar charts. The numerical summaries will provide a general overview of the center and spread/variation among the memory tests. The side-by-side boxplots will allow us to visualize the central tendencies and the spread of the memory tests among each group. Furthermore, histograms will allow us to analyze the shape of the results of the memory tests. The qualitative response variable will be visualized and analyzed through the stacked bar chart as it allows us to make comparisons across each group.*

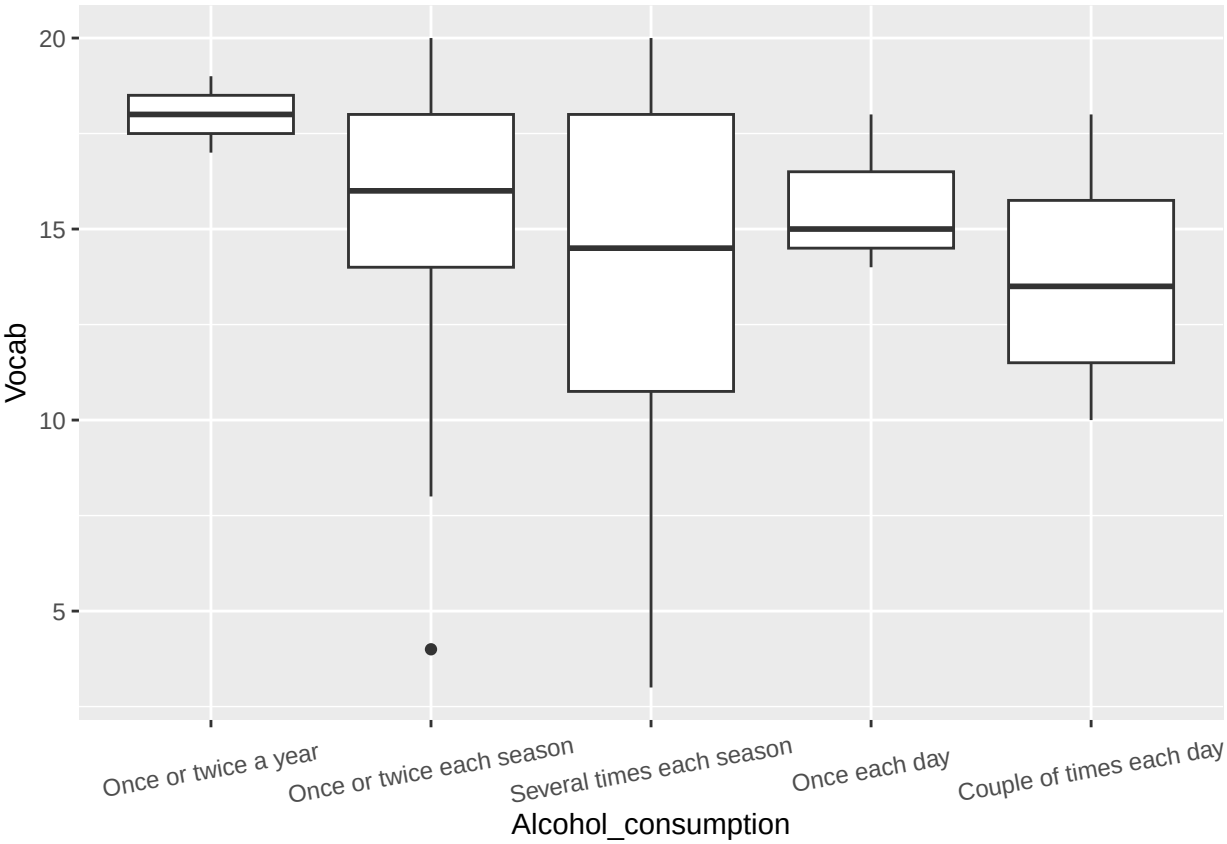
```
# side by side histograms

# cards
data %>%
  # order by increasing consumption
  mutate(Alcohol_consumption = factor(Alcohol_consumption, levels=c("Once or twice a year", "Once or twice eac
  ggplot(aes(x = Alcohol_consumption, y = Cards)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 10, vjust = 0.5))
```




```
# vocab
data %>%
  # order by increasing consumption
  mutate(Alcohol_consumption = factor(Alcohol_consumption, levels=c("Once or twice a year", "Once or twice eac
```

```
  ggplot(aes(x = Alcohol_consumption, y = Vocab)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 10, vjust = 0.5))
```

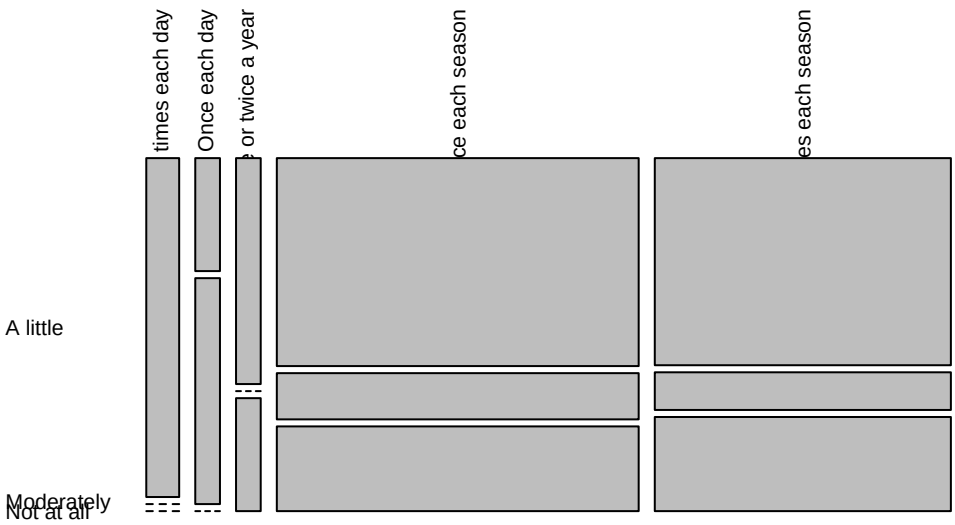



```
# mosaic of forgetfulness (totally unreadable)
# create two-way table
mostable <- table(data$Alcohol_consumption, data$Forgetful)
mostable
```

```
##
##                            A little Moderately Not at all
##   Couple of times each day         4          0          0
##   Once each day                    1          2          0
##   Once or twice a year             2          0          1
##   Once or twice each season       27          6         11
##   Several times each season       22          4         10
```

```
# create mosaic plot
mosaicplot(mostable, main = "Self-reported Forgetfulness by Alcohol Consumption Frequency", las=2)
```

Self−reported Forgetfulness by Alcohol Consumption Frequency




```
# replace with ggplot
```

Hypothesis tests

```
anova_cards <- aov(data$Cards ~ data$Alcohol_consumption)
summary(anova_cards)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## data$Alcohol_consumption  4   11.3   2.819   0.673  0.612
## Residuals                85  355.9   4.187
```

```
anova_vocab <- aov(data$Vocab ~ data$Alcohol_consumption)
summary(anova_vocab)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## data$Alcohol_consumption  4  105.1   26.27   1.715  0.154
## Residuals                85 1301.8   15.32
```

```
chisq.test(data$Alcohol_consumption, data$Forgetful)
```

```
## Warning in chisq.test(data$Alcohol_consumption, data$Forgetful): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$Alcohol_consumption and data$Forgetful
## X-squared = 10.789, df = 8, p-value = 0.214
```

Results

Conclusions

Discussion

References

Appendix