Stat 6021: Addressing Guided Question Set 1

Tom Lever

08/24/22

Download the dataset `students.txt` from Collab. The dataset contains information on students taking an introductory statistics class at a large public university in the early 2000's. The columns of data are:

```
* `Student`: ID number on survey
* `Gender`: gender of student (male / female)
* `Smoke`: whether the student smokes (yes / no)
* `Marijuan`: whether the student smokes marijuana (yes / no)
* `DrivDrnk`: whether the student has ever driven while drunk (yes / no)
* `GPA`: student's current GPA
* `PartyNum`: number of days per month the student parties
* `DaysBeer`: number of days per month the student has at least two alcoholic drinks
* `StudyHrs`: number of hours spent studying per week
```

For the questions below, you may use either the traditional / Base R approach or the `dplyr` approach (or even a combination of both approaches).

1. Looking at the variables above, is there a variable that will definitely not be part of any meaningful analysis? If yes, which one? Remove this variable from your data frame.

   Meaningful analyses would likely anonymize data by removing any student ID numbers.

   ```
   library(dplyr)
   students_dataframe <- read.table("students.txt", header=TRUE)
   head(students_dataframe, n = 3)
   ```

   ```
   ##   Student Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
   ## 1       1 female    No      Yes      Yes 3.40        4        6        7
   ## 2       2 female    No       No       No 3.45        4        0       20
   ## 3       3   male    No       No      Yes 3.89        9        4       30
   ```

   ```
   anonymized_students_dataframe <- students_dataframe%>%select(-Student)
   head(anonymized_students_dataframe, n = 3)
   ```

   ```
   ##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
   ## 1 female    No      Yes      Yes 3.40        4        6        7
   ## 2 female    No       No       No 3.45        4        0       20
   ## 3   male    No       No      Yes 3.89        9        4       30
   ```

2. How many students are there in this dataset?

   ```
   nrow(students_dataframe)
   ```

   ```
   ## [1] 249
   ```

   There are 249 students in the students dataset. The header row of `students_dataframe` is not considered in this determination.

3. How many students have a missing entry in at least one of the columns?

```r
number_of_students_with_missing_datum <-
    nrow(students_dataframe[!complete.cases(students_dataframe),])
cat(
    "There are ",
    number_of_students_with_missing_datum,
    " students with a missing datum."
)
```

```
## There are  12  students with a missing datum.
```

4. Report the median values of the numeric variables other than `Student`.

   The numeric variables other than `Student` are `GPA`, `PartyNum`, `DaysBeer`, and `StudyHrs`.

```r
median_values <- students_dataframe%>%
    summarize(
        across(c(GPA, PartyNum, DaysBeer, StudyHrs), ~median(.x, na.rm = TRUE))
    )
median_values
```

```
##   GPA PartyNum DaysBeer StudyHrs
## 1 3.2        8        8       14
```

```r
median_values <- t(
    students_dataframe%>%
        summarize(
            median_GPA = median(GPA, na.rm = TRUE),
            median_number_of_days_with_party_per_month =
                median(PartyNum, na.rm = TRUE),
            median_number_of_days_with_alcohol_per_month =
                median(DaysBeer, na.rm = TRUE),
            median_hours_study_per_week = median(StudyHrs, na.rm = TRUE)
        )
)
colnames(median_values) <- "median"
median_values
```

```
##                                              median
## median_GPA                                      3.2
## median_number_of_days_with_party_per_month      8.0
## median_number_of_days_with_alcohol_per_month    8.0
## median_hours_study_per_week                    14.0
```

5. Report the mean and standard deviation of `StudyHrs` for female and male students.

```r
students_dataframe%>%
    group_by(Gender)%>%
        summarize(
            mean_hours_study_per_week = mean(StudyHrs, na.rm = TRUE),
            standard_deviation = sd(StudyHrs, na.rm = TRUE)
        )
```

```
## # A tibble: 2 x 3
##   Gender mean_hours_study_per_week standard_deviation
##   <chr>                      <dbl>              <dbl>
## 1 female                      15.4               8.97
## 2 male                        14.7              10.2
```

6. Construct a 95-percent confidence interval for the mean `StudyHrs` for female students, and another 95-percent confidence interval for the mean `StudyHrs` for male students. Based on these intervals, do we have evidence that the mean `StudyHrs` is different between female and male students? **Hint:** use the `table()` function (base R) or the `count()` function from the `dplyr` package to obtain the sample sizes of female and male students.

   See Tom Lever's R Package.

```
library(TomLeversRPackage)
female_students_dataframe <- students_dataframe%>%filter(Gender=="female")
male_students_dataframe <- students_dataframe%>%filter(Gender=="male")
hours_study_per_week_for_female_students <-
    female_students_dataframe%>%pull(StudyHrs)
hours_study_per_week_for_male_students <- male_students_dataframe%>%pull(StudyHrs)
constructConfidenceIntervalForPopulationMean(
    hours_study_per_week_for_female_students,
    0.05
)
```

```
## [1] 13.93409 16.87970
```

```
constructConfidenceIntervalForPopulationMean(
    hours_study_per_week_for_male_students,
    0.05
)
```

```
## [1] 12.71850 16.68535
```

```
constructConfidenceIntervalForDifferenceBetweenTwoPopulationMeans(
    hours_study_per_week_for_female_students,
    hours_study_per_week_for_male_students,
    0.05
)
```

```
## [1] -1.700220  3.110167
```

   Since there is overlap between the confidence interval for the population mean hours study per week for female students and the confidence interval for the population mean hours study per week for male students, we conservatively state that we do not have sufficient evidence to conclude that there is adifference between mean study hours per week for female and male student populations.

   Since 0 is in the above confidence interval, a difference of 0 between mean study hours per week for female and male student populations is within the margin of error of the confidence interval for 95 percent of samples. We do not have sufficient evidence to conclude that there is a difference between mean study hours per week for female and male student populations.

7. Compare the median `StudyHrs` across genders and `Smoke`.

```
contigency_table <- tapply(
    students_dataframe%>%pull(StudyHrs),
    list(students_dataframe%>%pull(Gender), students_dataframe%>%pull(Smoke)),
    median,
    na.rm = TRUE,
    simplify = TRUE,
    default = -1
)
contigency_table
```

```
##        No Yes
## female 15  10
## male   12  14
```

```
chisq.test(contigency_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contigency_table
## X-squared = 0.50373, df = 1, p-value = 0.4779
```

Since median hours study per week for female students going left to right decreases, and median hours study per week for male students going left to right increases, there may be an association between `Gender` and `Smoke`.

Given a significance level 0.05, since the above probability 0.478 is greater than the significance level, we fail to reject the null hypothesis of the Pearson's Chi-squared test of independence, which states that there is no association between the row variable `Gender` and the column variable `Smoke`. There may be an association between `Gender` and `Smoke`.

8. Create a new variable called `PartyAnimal`, which takes on the value "yes" if the student parties a lot (more than 8 days per month) (i.e., `PartyNum` > 8), and "no" otherwise.

```
library(dplyr)
PartyNum <- students_dataframe%>%select(PartyNum)
PartyAnimal <- ifelse(PartyNum > 8, "yes", "no")
colnames(PartyAnimal) <- "PartyAnimal"
head(bind_cols(PartyNum, PartyAnimal), n = 7)
```

```
##   PartyNum PartyAnimal
## 1        4          no
## 2        4          no
## 3        9         yes
## 4        6          no
## 5       10         yes
## 6        2          no
## 7        8          no
```

9. Create a new variable called `GPA.cat`, which takes on the following values.

   - "low" if GPA is less than 3.00 (less than or equal to 2.99)
   - "moderate" if GPA is at least 3.00 and less than 3.50 (less than or equal to 3.49)
   - "high" if GPA is at least 3.50

```
GPA <- c(2.99, 3.00, 3.01, 3.49, 3.50, 3.51)
GPA.cat <-
    data.frame(
        GPA.cat =
            cut(
                GPA,
                #breaks = c(-Inf, 2.99, 3.49, Inf),
                breaks = c(-Inf, 3.00, 3.50, Inf),
                right = FALSE,
                labels = c("low", "moderate", "high")
            )
    )
bind_cols(data.frame(GPA = GPA), GPA.cat)
```

```
##    GPA  GPA.cat
## 1 2.99      low
## 2 3.00 moderate
## 3 3.01 moderate
## 4 3.49 moderate
## 5 3.50     high
## 6 3.51     high
```

```
GPA <- students_dataframe%>%pull(GPA)
GPA.cat <-
    data.frame(
        GPA.cat =
            cut(
                GPA,
                breaks = c(-Inf, 3.00, 3.50, Inf),
                right = FALSE,
                labels = c("low", "moderate", "high")
            )
    )
head(bind_cols(data.frame(GPA = GPA), GPA.cat), n = 5)
```

```
##    GPA  GPA.cat
## 1 3.40 moderate
## 2 3.45 moderate
## 3 3.89     high
## 4 3.75     high
## 5 2.30      low
```

10. Add the variables `PartyAnimal` and `GPA.cat` to the redacted data frame from part [1], and export it to a `.csv` file. Name the file `new_students.csv`. We will be using this data file for the next module.

```
new_students_dataframe <- bind_cols(students_dataframe, PartyAnimal, GPA.cat)
head(new_students_dataframe, n = 3)
```

```
##   Student Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
## 1       1 female    No      Yes      Yes 3.40        4        6        7
## 2       2 female    No       No       No 3.45        4        0       20
## 3       3   male    No       No      Yes 3.89        9        4       30
##   PartyAnimal  GPA.cat
## 1          no moderate
## 2          no moderate
## 3         yes     high
```

```
write.csv(new_students_dataframe, "new_students.csv", row.names = FALSE)

new_anonymized_students_dataframe <-
    bind_cols(anonymized_students_dataframe, PartyAnimal, GPA.cat)
head(new_anonymized_students_dataframe, n = 3)
```

```
##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs PartyAnimal
## 1 female    No      Yes      Yes 3.40        4        6        7          no
## 2 female    No       No       No 3.45        4        0       20          no
## 3   male    No       No      Yes 3.89        9        4       30         yes
##    GPA.cat
## 1 moderate
## 2 moderate
## 3     high
```

```
write.csv(
    new_anonymized_students_dataframe,
    "new_anonymized_students.csv",
    row.names = FALSE
)
```

11. Suppose we want to focus on students who have low (i.e., below 3.00) GPA's, party a lot (i.e., more than 8 days per month), and study little (i.e., less than 15 hours per week). Create a data frame that contains these students. How many such students are there?

```
dataframe_for_students_who_have_low_GPAs_party_a_lot_and_study_little <-
    students_dataframe%>%
        filter(GPA < 3.00, PartyNum > 8, StudyHrs < 15)
head(dataframe_for_students_who_have_low_GPAs_party_a_lot_and_study_little, n = 3)
```

```
##   Student Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
## 1       5   male   Yes      Yes      Yes 2.30       10       15       14
## 2       9 female    No      Yes      Yes 1.87       16       20        6
## 3      18 female    No      Yes      Yes 2.70        9        8        9
```

```
cat(
    "There are ",
    nrow(dataframe_for_students_who_have_low_GPAs_party_a_lot_and_study_little),
    " students in this dataset.\n",
    "The header row of students_dataframe is not considered in this determination."
)
```

```
## There are  29  students in this dataset.
##  The header row of students_dataframe is not considered in this determination.
```