Student Analysis

Jason Ballantyne

03/11/2021

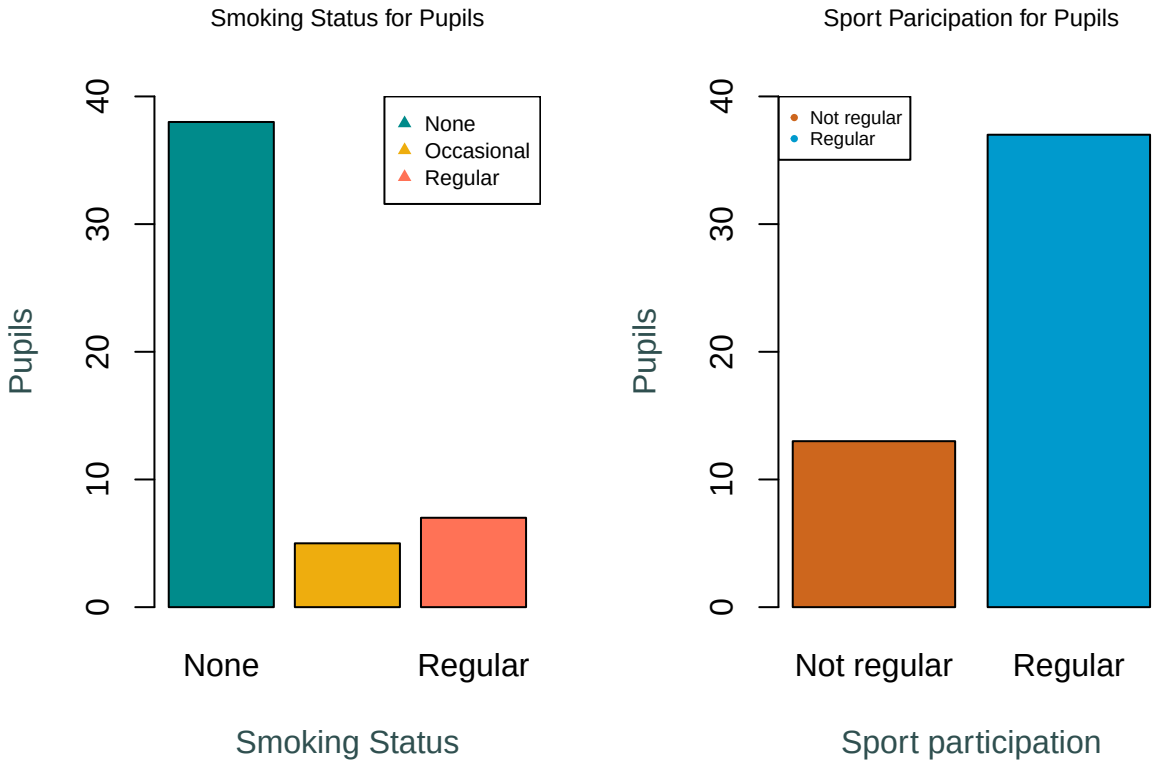
1.

```
# Load in the data
Substance_df <- read.delim(file = "C:\\Users\\jason\\OneDrive - University College Dublin\\Documents\\M

# Converting alcohol column to ordered factor with appropriate labels
Substance_df$alcohol <- factor(Substance_df$alcohol,
                               ordered = TRUE,
                               levels = c(1, 2, 3, 4, 5),
                               labels = c("None", "Once or twice a year",
                                          "Once a month", "Once a week",
                                          "More than once a week"))

# Converting drugs column to ordered factor with appropriate labels
Substance_df$drugs <- factor(Substance_df$drugs,
                             ordered = TRUE,
                             levels = c(1, 2, 3, 4),
                             labels = c("None", "Tried once",
                                        "Occasional", "Regular"))

# Converting smoke column to ordered factor with appropriate labels
Substance_df$smoke <- factor(Substance_df$smoke,
                             ordered = TRUE,
                             levels = c(1, 2, 3),
                             labels = c("None", "Occasional", "Regular"))

# Converting sport column to ordered factor with appropriate labels
Substance_df$sport <- factor(Substance_df$sport,
                             ordered = TRUE, levels = c(1, 2),
                             labels = c("Not regular", "Regular"))

# Display the structure of the dataset
str(Substance_df)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ alcohol: Ord.factor w/ 5 levels "None"<"Once or twice a year"<..: 3 2 2 2 3 4 4 4 2 4 ...
##  $ drugs  : Ord.factor w/ 4 levels "None"<"Tried once"<..: 1 2 1 1 1 1 3 3 1 1 ...
##  $ smoke  : Ord.factor w/ 3 levels "None"<"Occasional"<..: 2 3 1 1 1 1 1 3 1 1 ...
##  $ sport  : Ord.factor w/ 2 levels "Not regular"<..: 2 1 1 2 2 2 1 2 2 2 ...
```

2.

```
# Creating two graphs illustrating the variables smoke and sport using base R
# Both graphs contain labels, colours and legends
# For comparison, the two plots will be put next to
# each other on the same page using the par function

# Putting plots next to each other on the same page
par(mfrow=c(1,2))

# Graph 1 - Using the smoke variable
plot(Substance_df$smoke,
     ylim = c(0,40),
     xlab = "Smoking Status",
     ylab = "Pupils",
     font.lab = 17,
     col.lab = "darkslategrey",
     main = "Smoking Status for Pupils",
     font.main=10,
     col.main = "black",
     pch = 17,
     cex.main = 0.75,
     col=ifelse(Substance_df$smoke == "None","coral1",
                ifelse(Substance_df$smoke == "Occasional","cyan4",
                       ifelse(Substance_df$smoke ==  "Regular",
                              "darkgoldenrod2", "grey"))))


# legend for smoke graph
legend("topright", c("None","Occasional", "Regular"),
       col = c("cyan4","darkgoldenrod2", "coral1"), pch=17, cex=0.70)


# Graph 2 - Using the sport variable

plot(Substance_df$sport,
     ylim = c(0,40),
     xlab = "Sport participation",
     ylab = "Pupils",
     font.lab = 15,
     col.lab = "darkslategrey",
     main = "Sport Paricipation for Pupils",
     font.main=10,
     col.main = "black",
     pch = 20,
     cex.main = 0.75,
     col=ifelse(Substance_df$sport == "Not regular","deepskyblue3",
                ifelse(Substance_df$sport ==  "Regular",
                              "chocolate3", "grey")))

# legend for sport graph
legend("topleft", c("Not regular","Regular"),
       col = c("chocolate3","deepskyblue3"), pch=20, cex=0.55)
```



Graph 1 - Smoke Variable

From our first graph on the smoke variable, it is clear that there is a huge disparity between pupils who don't smoke and our remaining two smoking status options which are occasional smokers and regular smokers. An overwhelming majority of pupils prefer not to smoke at all. The second most common status amoung pupils is regular (more than once per week) smokers. This is closely followed by occasional smokers.

Graph 2 - Sport Variable

From our second graph on the sport variable, there is once again, a huge disparity between our two options, regular and non regular sport players. Over twice the number of pupils regularly participate in sport as opposed to those who do not. We see a major drop in numbers with respect to pupils who do not regularly participate in spot based on our graph.

3.

(i)

```
# Finding the proportion of pupils who smoke at least occasionally
# We must calculate the pupils who smoke occasionally & regularly

smoke_cout <- table(Substance_df$smoke)
```

```r
# Full table of proportions for pupils who smoke
smoke_proportion <- (smoke_cout/length(Substance_df$smoke))

# Calculating the proportion of pupils who smoke at least occasionally
smokers <- smoke_proportion["Occasional"] + smoke_proportion["Regular"]
smokers
```

```
## Occasional
##       0.24
```

- Based on the output above, the proportion of pupils who smoke **at least** occasionally is 0.24.

(ii)

```r
# Finding the proportion of pupils who regularly practice sport and
# smoke at least occasionally

# Summing the number of regular sport pupils and occasional or regular smokers
sport_smokers <- sum(Substance_df$sport == "Regular" &
                       (Substance_df$smoke == "Occasional" |
                          Substance_df$smoke == "Regular"))

# Getting the total number of pupils
total_pupils <- nrow(Substance_df)

#Calculating the proportion of pupils who regularly practice sport and smoke
# at least occasionally and printing the result
sport_smokers_proportion <- (sport_smokers/total_pupils)
sport_smokers_proportion
```

```
## [1] 0.18
```

- Based on the output above, the proportion of pupils who regularly practice sport and smoke **at least** occasionally is 0.18.

4.

```r
# Turn object into an S3 class called s50survey
# Summary method that will show the proportion of students
# for every level of each variable

new_s50survey <- function(lst){
    structure(lst, class = "s50survey" )
}

# Summary method that shows the proportion of students for
# every level of each variable
summary.s50survey <- function(survey){
```

```
    lapply(survey, \(x) prop.table(table(x, dnn = "")))
}

# Testing my function on the s50_1995.txt data
testing_data <- Substance_df

# Turning the testing data into an S3 class called s50survey
s50survey <- new_s50survey(testing_data)

# Passing the s50survey through our summary method to show proportion
# for every level of each variable
summary(s50survey)
```

```
## $alcohol
##
##                  None  Once or twice a year          Once a month
##                  0.10                  0.32                  0.24
##           Once a week More than once a week
##                  0.28                  0.06
##
## $drugs
##
##       None Tried once Occasional    Regular
##       0.72       0.12       0.14       0.02
##
## $smoke
##
##       None Occasional    Regular
##       0.76       0.10       0.14
##
## $sport
##
## Not regular     Regular
##        0.26        0.74
```

5.

```
# Finding the proportion of pupils who did not use cannabis
drugs_cout <- table(Substance_df$drugs)

# Full table of proportions for drug use amoung pupils
drugs_proportion <- (drugs_cout/length(Substance_df$drugs))

# Calculating the proportion of pupils who did not use cannabis
no_drugs <- drugs_proportion["None"]
no_drugs
```

```
## None
## 0.72
```

- Based on the output above, the proportion of pupils who did not use cannabis is 0.72.

6.

```
# Reading in the s50_1997.txt
# Converting each column to an ordered factor with appropriate labels
# Assigning the class s50survey to this new dataset
# Testing the summary S3 method on this new dataset


# Load in the data
Substance_1997_df <- read.delim(file = "C:\\Users\\jason\\OneDrive - University College Dublin\\Documen

# Converting alcohol column to ordered factor with appropriate labels
Substance_1997_df$alcohol <- factor(Substance_1997_df$alcohol,
                                    ordered = TRUE, levels = c(1, 2, 3, 4, 5),
                                    labels = c("None",
                                               "Once or twice a year",
                                               "Once a month",
                                               "Once a week",
                                               "More than once a week"))

# Converting drugs column to ordered factor with appropriate labels
Substance_1997_df$drugs <- factor(Substance_1997_df$drugs,
                                  ordered = TRUE, levels = c(1, 2, 3, 4),
                                  labels = c("None", "Tried once",
                                             "Occasional", "Regular"))

# Converting smoke column to ordered factor with appropriate labels
Substance_1997_df$smoke <- factor(Substance_1997_df$smoke,
                                  ordered = TRUE, levels = c(1, 2, 3),
                                  labels = c("None", "Occasional", "Regular"))

# Converting sport column to ordered factor with appropriate labels
Substance_1997_df$sport <- factor(Substance_1997_df$sport,
                                  ordered = TRUE, levels = c(1, 2),
                                  labels = c("Not regular", "Regular"))


# Function to turn the dataset to an S3 class called s50survey
new_s50survey <- function(lst){
    structure(lst, class = "s50survey" )
}

# Summary method that shows the proportion of students
# for every level of each variable
summary.s50survey <- function(survey){
    lapply(survey, \(x) prop.table(table(x, dnn = "")))
}

new_testing_data = Substance_1997_df

# Turning the new dataset into an S3 class
s50survey <- new_s50survey(new_testing_data)
```

```
# Passing the s50survey through our summary
# method to show proportion for every level of each variable
summary(s50survey)
```

```
## $alcohol
##
##                  None  Once or twice a year          Once a month
##                  0.02                  0.18                  0.34
##           Once a week More than once a week
##                  0.34                  0.12
##
## $drugs
##
##       None Tried once Occasional    Regular
##       0.52       0.14       0.34       0.00
##
## $smoke
##
##       None Occasional    Regular
##       0.62       0.04       0.34
##
## $sport
##
## Not regular     Regular
##        0.62        0.38
```

7.

```
# Checking if the proportion of students practising
# sport regularly increased or decreased with respect to the 1995 data

# Summing the number of regular sport pupils in 1997
sport_1997 <- sum(Substance_1997_df$sport == "Regular")

# Getting the total number of pupils in 1997
total_1997 <- nrow(Substance_1997_df)

# Calculating the proportion of pupils who regularly practice sport in 1997
sport_players_1997 <- (sport_1997/total_1997)



# Summing the number of regular sport pupils in 1995
sport_1995 <- sum(Substance_df$sport == "Regular")

# Getting the total number of pupils in 1995
total_1995 <- nrow(Substance_df)

# Calculating the proportion of pupils who regularly practice sport in 1995
sport_players_1995 <- (sport_1995/total_1995)
```

```r
# Creating a dataframe to illustrate the difference
# in proportions between 1995 and 1997
RegularSports_df = data.frame(
        col1 = (sport_players_1995),
        col2 = (sport_players_1997)
)
colnames(RegularSports_df) <- c('1995 Regular Sports','1997 Regular Sports')

RegularSports_df
```

```
##   1995 Regular Sports 1997 Regular Sports
## 1                0.74                0.38
```

- The proportion of students practising sport regularly **decreased** with respect to the 1995 data.
- The proportion of students practising sport regularly in 1995 was 0.74
- The proportion of students practising sport regularly in 1997 was 0.38