Lecture 0 Handout: Introduction to R

Statistics 139 Team

September 05, 2023

Topics

- R, RStudio
- Data Collection, Data Wrangling, and Exploratory Data Analysis
- Hypothesis Testing and Confidence Interval Review

The material in this lab corresponds to Lecture 0 notes.

Be sure to fill out the following survey (5 questions) before continuing in this lab:

https://bit.ly/lecturesurvey0

A survey was conducted of Stat 139 students and measured 5 variables (responses will be saved as 'stat139survey.csv'):

- heartrate: beats per minute
- exercise: the number of hours of vigorous exercise in a typical week
- coffee: an indicator variable measuring whether respondent drank coffee that day
- gender: male, female, non-binary, or other
- class: freshman, sophomore, junior, senior, grad, or other

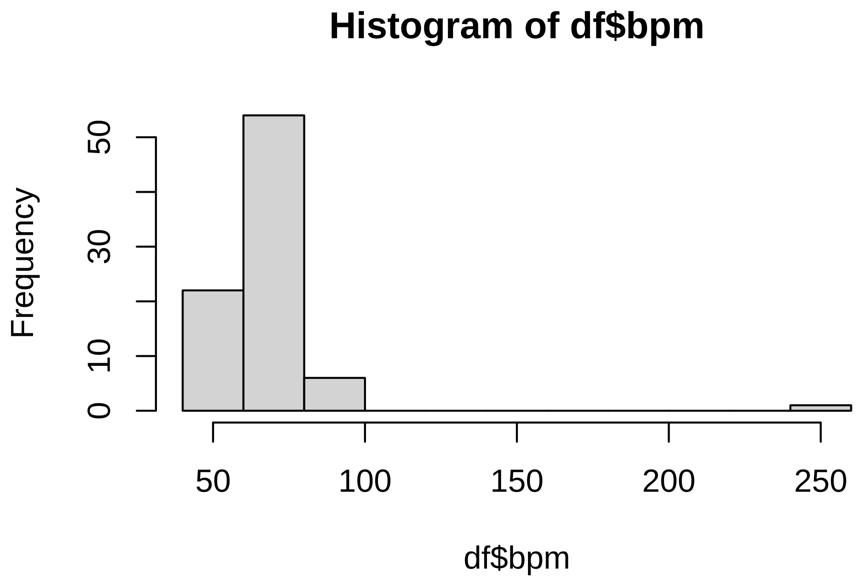
```
library(ggplot2)
library(data.table)
```

Question 1.

a) We are mainly interested in measuring heartrate from this survey. Perform a little exploratory data analysis (EDA) on this variable. Provide a basic visual and calculate summary statistics.

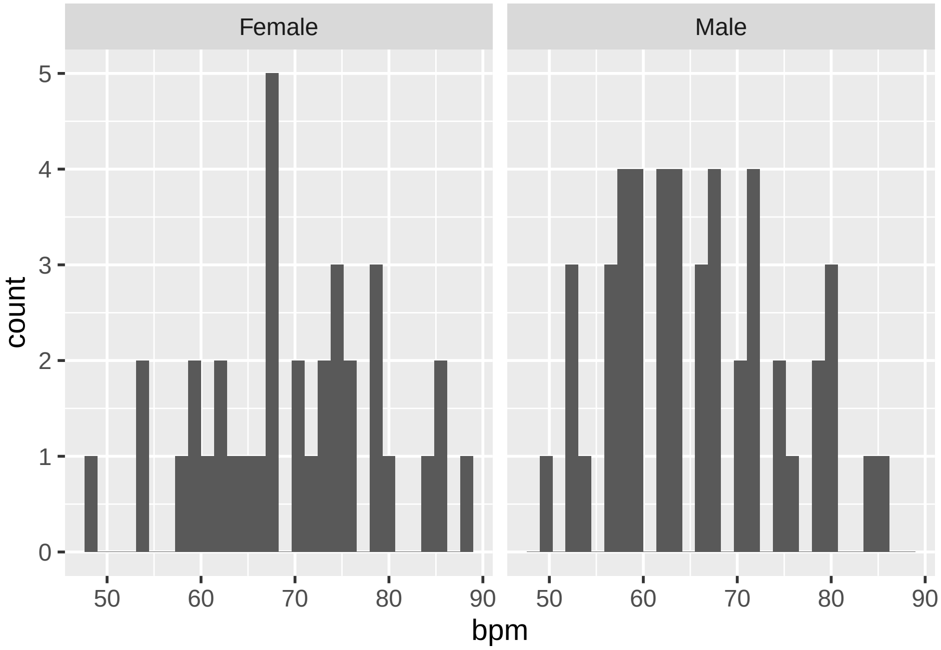
```
#read in the data
surveydata = read.csv("./data/stat139survey.csv")

# select for relevant columns
df <- surveydata[, 3:7]
```

```
# rename variables
colnames(df) <- c("bpm", "exercise", "coffee", "gender", "class")

# change to datatable
df <- as.data.table(df)

# provide a visual and calculate summary statistics
hist(df$bpm)
```



```
# faceted histogram by gender
ggplot(df[df$bpm < 200, ], aes(bpm)) +
  geom_histogram() +
  facet_wrap(~gender)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
# calculate mean
mean(df$bpm)
```

```
## [1] 69.6747
```

```r
# look at structure
str(df)
```

```
## Classes ’data.table’ and ’data.frame’:   83 obs. of  5 variables:
##  $ bpm     : int  80 80 73 72 64 68 52 74 86 68 ...
##  $ exercise: chr  "0" "0.5" "7" "5" ...
##  $ coffee  : chr  "No" "Yes" "No" "No" ...
##  $ gender  : chr  "Male" "Female" "Female" "Male" ...
##  $ class   : chr  "Junior" "Grad Student" "Junior" "Junior" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
# look at values of exercise var
table(df$exercise)
```

```
##
##  -2   0 0.5   1  10  12  14  15  18   2  20   3   4 4-6 4.5   5   6   7   8   9
##   1   5   1   8   3   3   1   1   2   7   1   9   7   1   1   8  11   7   4   2
```

```r
# wrangle inappropriate values
df[exercise == "4-6", exercise := "5"]
df <- df[exercise != "-2"]
```

```
# looks good!
table(df$exercise)
```

```
## 
##   0 0.5   1  10  12  14  15  18   2  20   3   4 4.5   5   6   7   8   9 
##   5   1   8   3   3   1   1   2   7   1   9   7   1   9  11   7   4   2
```

```
# recode to numeric
df[, exercise := as.numeric(exercise)]
```

Recall the classic formula for the confidence interval for a population mean:

$$\bar{x} \pm t^* \left(\frac{s}{\sqrt{n}} \right)$$

b) Use this data set to provide a 95% confidence interval for the true mean heartrate of all Harvard students.

```
# calculate sample statistics: sample size, sample mean, and sample sd
sample_size <- nrow(df)
sample_mean <- mean(df$bpm)
sample_sd <- sd(df$bpm)

# pull off the correct quantile from the t-distribution
# qt(p, df)
p <- 0.975
degree_f <- sample_size - 1
quantile <- qt(p, degree_f)

# calculate the appropriate confidence interval
sample_mean - quantile * sample_sd/(sqrt(sample_size))
sample_mean + quantile * sample_sd/(sqrt(sample_size))
```

c) Interpret the confidence interval you calculated in the previous part.

d) The true mean resting heartrate for adults in the US is reported to be 70 beats per minute. How does the calculated confidence interval compare to the US population? What can you conclude?

The calculated confidence interval aligns with what is reported to be the avg heartrate for adults in the US

e) What assumptions are needed for the confidence interval to be exact? To be approximately correct?

f) Do you trust the inferences above? Why or why not?

Question 2. Harvard reports that 49.5% of undergraduate students at the college are male.

a) What proportion of undergraduate students in this survey are male?

```r
mean(df$gender == "Male")
```

```
## [1] 0.5731707
```

b) Write down the formula for the confidence interval for a population proportion. Calculate this interval from the data.

c) What is a reasonable target population for this survey? What can you conclude from the confidence interval above?

d) What assumptions are needed for the confidence interval in part (b) to be reasonable? Is it ever exact? Why or why not?

Question 3. What other interesting questions could be answered with this data set? Perform some exploratory analyses (both visually and statistically) to investigate these questions.