HW1

Kaylee Vo

2025-06-26

Make sure you put your name in the author and date above.

16 points total for this section.

[0 points] First, makes sure you set your "HSS" folder as your working document! (refer to Lecture 1 slides 68-72) Then, load the NHANES dataset - we have provided the code for you here to run - if your working directory is not set up correctly, this will have errors.

```
load("NHANES.RData")
```

Creating a new dataset with your study base with subset(...)

[2 points] Create an object called NHANES_2 by doing subset(...) on the original NHANES dataset and capturing only those meeting the inclusion criteria according to the flowchart above.

Beware that this study is 21 to 40 years old, inclusive - make sure that you use the greater/less than and/or equals to (<, >, <=, >=) effectively.

```
#Run the code for the subset below:
NHANES_2 <- subset(NHANES, age >= 21 & age <= 40 & !is.na(alcohol_amount))
```

Now, please CHECK that your subsetting was successful

[2 points] (i) Check the age range - first, show that the original NHANES includes ages 0-80; then, show that NHANES_2 indeed includes only ages 21-40.

[2 points] (ii) Check for missing data - first, show the number of missing data for alcohol_amount in the original NHANES; then, show that NHANES_2 does not have missing data.

```
#Code for (i) - 2 lines:
range(NHANES$age)
```

```
## [1]  0 80
```

```
range(NHANES_2$age)
```

```
## [1] 21 40
```

```
#Code for (ii) - 2 lines:
sum(is.na(NHANES$alcohol_amount))
```

```
## [1] 4540
```

```r
sum(is.na(NHANES_2$alcohol_amount))
```

```
## [1] 0
```

Creating a new variable toevaluate excess alcohol consumption (demo)

[0 points] Now, you will create a new variable called HiAlcohol in the NHANES_2 dataset, where you will designate those who consume 5 or more (>=5) drinks on average, on days that they drank alcoholic beverages, as "Yes" and those who consume less than 5 (<5) as "No" - beware of the quotation mark.

Step 1: You will use the dataset$variable <- ifelse(....) code to create the new variable. Step 2: Then, you will make the HiAlcohol variable into a factor type. Step 3: Then, you will CHECK that it became a factor.

We are showing you step-by-step how to do this now, but going forward, you will be expected to do something like this yourself. Please make sure you are comfortable running this code - we will also cover this during class.

```r
#We have provided the code below for you - please click on the green arrow to run it before proceeding.

NHANES_2$hi_alcohol <- ifelse(
  NHANES_2$alcohol_amount >=5, "Yes",
    ifelse(NHANES_2$alcohol_amount <5, "No", NA)
)

data.class(NHANES_2$hi_alcohol)
```

```
## [1] "character"
```

```r
NHANES_2$hi_alcohol <- factor(NHANES_2$hi_alcohol)

data.class(NHANES_2$hi_alcohol)
```

```
## [1] "factor"
```

```r
#If this ran correctly, you will have no error messages and will see this below:
#[1] "character"
#[1] "factor"
```

Excess alcohol consumption in the total study base

[2 points] (i) Use the table(dataset$var, exclude=F) command to find the number of those with Yes vs. No to hi_alcohol.

[2 points] (ii) Using the output, calculate the prevalence of excessive alcohol consumption overall. Show your calculation [use R as a giant calculator]. Recall that the numerator is the number of Yes, the denominator is the total number of subjects in the study (i.e. the number of observations in your NHANES_2 dataset).

```r
#Code for (i)
alc_table = table(NHANES_2$hi_alcohol, exclude=F)
alc_table
```

```
##
##   No  Yes
## 1630  361
```

```
#Code for (ii)
alc_prevalence = alc_table['Yes'] / nrow(NHANES_2)
names(alc_prevalence) = NULL
alc_prevalence
```

```
## [1] 0.1813159
```

Excess alcohol consumption by sex

[2 points] (i) Use the table(dataset$var1$, $dataset$var2, exclude=F) command to find the number of those with Yes vs. No to hi_alcohol, by sex (NHANES only provides Female vs. Male dichotomous self-reported sex).

[2 points] (ii) Using the output, calculate the prevalence of hi_alcohol among females. You would calculate the number of Female hi_alcohol=Yes divided by the total number of Females (hi_alcohol=Yes + hi_alcohol=No groups would equal to the total) [use R as a giant calculator].

[2 points] (iii) Then, calculate the prevalence of hi_alcohol among males. You would calculate the number of Male hi_alcohol=Yes divided by the total number of Males (hi_alcohol=Yes + hi_alcohol=No groups would equal to the total) [use R as a giant calculator].

```
#Code for (i):
alc_sex_table = table(NHANES_2$hi_alcohol, NHANES_2$sex, exclude=F)
alc_sex_table
```

```
##
##       Male Female
##   No   798    832
##   Yes  246    115
```

```
#Code for (ii):
fem_prevalence = alc_sex_table['Yes', 'Female'] / (alc_sex_table['Yes', 'Female'] + alc_sex_table['No',
fem_prevalence
```

```
## [1] 0.1214361
```

```
#Code for (iii):
male_prevalence = alc_sex_table['Yes', 'Male'] / (alc_sex_table['Yes', 'Male'] + alc_sex_table['No', 'Ma
male_prevalence
```

```
## [1] 0.2356322
```