STAT 630: Homework 1

Ly Que Nguyen

Due: September 7th, 2023 at 11:59pm

1. Listen to the following episode of Stats + Stories.

    a) What is the sampling unit in the American Housing Survey?

Housing unit

`b) What does HUD stand for?`

Housing and Urban Development

`c) How does the Census Bureau try to reduce respondent burden?`

Through the use of administrative data, property data should be public information, the county collects information about all the residences and the property in their area. For taxation purposes, they want to know the value of the land and the building on that land, to tax its residents. But to do that, they need detailed information. They need a number of bedrooms, they need your belt of things that impact their valuation of the housing value. And this data exists throughout the country. And it's public. And there are now data products out there that go out, put all this data to gather throughout the country to gather and deck one data file and it can be used.

`d) Describe the sampling process in a few sentences _in your own words_.`

The American Housing Survey is a long-running, comprehensive study of housing units and their occupants since 1973. It covers various aspects, including unit features, resident demographics, and housing costs. Sampling starts with the Census Bureau's Master Address File, which compiles US addresses. About 100,000 units are selected, enabling national and regional estimates for large cities and other statistical areas. Conducted biennially, residents provide details on their unit, like bedrooms, improvements, and nearby amenities. Demographic data is also collected. One notable feature of the American Housing Survey is its longitudinal nature. Notably, the survey revisits the same units every two years, offering insights into changing trends. Trends include more adult children and unrelated families living together. This reflects shifts in household composition. Additionally, the survey assesses housing quality and costs, providing crucial data for policy making.

`e) What would you like to know about? Write a research question that could be answered with the Americar`

How have trends in household composition, including an increase in adult children living at home and unrelated families sharing housing units, evolved over the past two decades?

2. Install the `openintro` package, by uncommenting the following code.

Reminder: you only have to do this once- like installing an app on your phone. After you run this line of code, either ***comment*** *it out using* ***#****, or just delete it.*

```
#install.packages("openintro")
```

After installing the R package for our book, load it, i.e., open the app!

```
library(openintro) # Load the package
```

Load in the `babies` dataset. Use the help file to learn more.

```r
data(babies) # Load the data
?babies # View the help file
```

View a summary of the dataset.

```r
dplyr::glimpse(babies) # Glimpse the dataset
```

```
## Rows: 1,236
## Columns: 8
## $ case      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ bwt       <int> 120, 113, 128, 123, 108, 136, 138, 132, 120, 143, 140, 144, ~
## $ gestation <int> 284, 282, 279, NA, 282, 286, 244, 245, 289, 299, 351, 282, 2~
## $ parity    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ age       <int> 27, 33, 28, 36, 23, 25, 33, 23, 25, 30, 27, 32, 23, 36, 30, ~
## $ height    <int> 62, 64, 64, 69, 67, 62, 62, 65, 62, 66, 68, 64, 63, 61, 63, ~
## $ weight    <int> 100, 135, 115, 190, 125, 93, 178, 140, 125, 136, 120, 124, 1~
## $ smoke     <int> 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, ~
```

```r
summary(babies) # View a summary of each column (variable)
```

```
##       case             bwt          gestation         parity
##  Min.   :   1.0   Min.   : 55.0   Min.   :148.0   Min.   :0.0000
##  1st Qu.: 309.8   1st Qu.:108.8   1st Qu.:272.0   1st Qu.:0.0000
##  Median : 618.5   Median :120.0   Median :280.0   Median :0.0000
##  Mean   : 618.5   Mean   :119.6   Mean   :279.3   Mean   :0.2549
##  3rd Qu.: 927.2   3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000
##  Max.   :1236.0   Max.   :176.0   Max.   :353.0   Max.   :1.0000
##                                   NA's   :13
##       age            height          weight          smoke
##  Min.   :15.00   Min.   :53.00   Min.   : 87.0   Min.   :0.0000
##  1st Qu.:23.00   1st Qu.:62.00   1st Qu.:114.8   1st Qu.:0.0000
##  Median :26.00   Median :64.00   Median :125.0   Median :0.0000
##  Mean   :27.26   Mean   :64.05   Mean   :128.6   Mean   :0.3948
##  3rd Qu.:31.00   3rd Qu.:66.00   3rd Qu.:139.0   3rd Qu.:1.0000
##  Max.   :45.00   Max.   :72.00   Max.   :250.0   Max.   :1.0000
##  NA's   :2       NA's   :22      NA's   :36      NA's   :10
```

a) What does each row in the table represent, i.e., what is the observational unit?

Each row provides data of 8 variables including the order of case, birth's weight, gestation, parity, age, height, weight and smoke status of each case/observation from 1 to 1238.

what is the observational unit?

Baby

b) How many participants were in the study?
1236 babies

c) All variables are coded as integers. Which variables should be recoded as *factors*?
Recode these variables in the code chunk below.

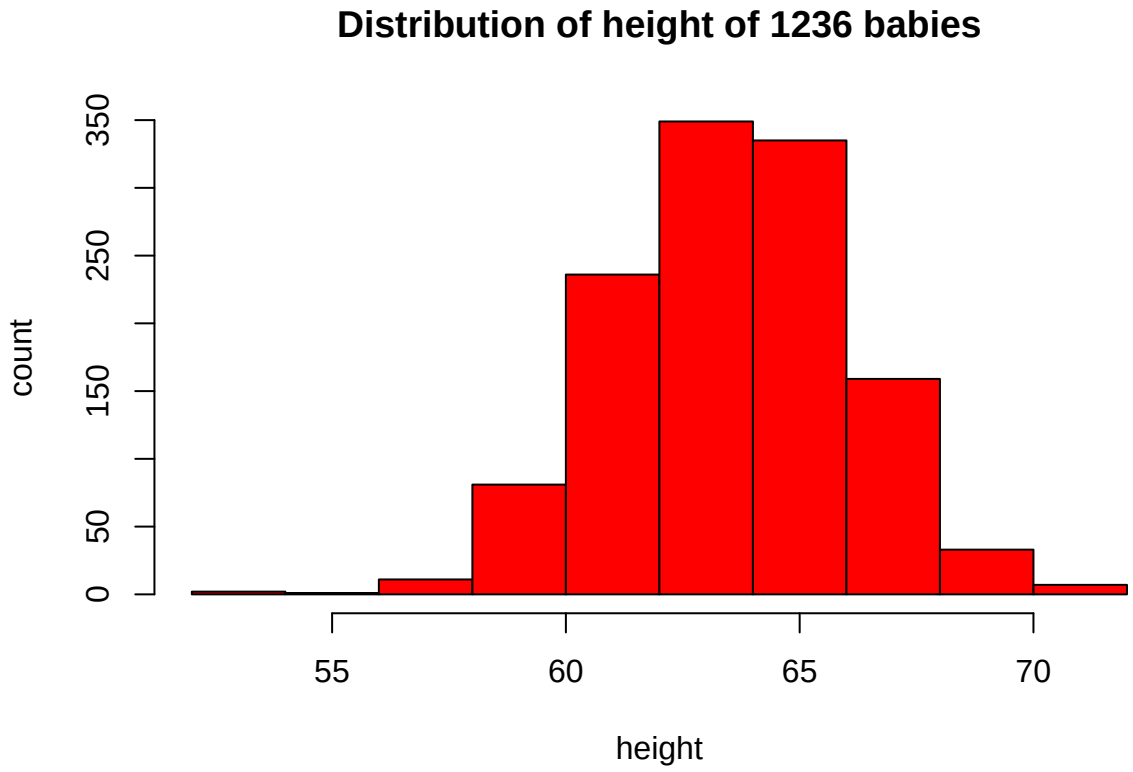
Smoke status should be recoded as factors.

```r
babies$smoke_f <- factor(babies$smoke,
                         labels = c("no", "yes"))
```

d) Create a plot to visualize one variable of your choice using `base R` functions. in the code chunk b

Make sure to add a title and relabel the x and y axes.

```
hist(babies$height,
     main = "Distribution of height of 1236 babies",
     breaks = 10,
     xlab = "height",
     ylab = "count",
     col = "red")
```



e) Create a plot to visualize the same variable from part (d) using `tidyverse` functions. in the code
Make sure to add a title and relabel the x and y axes.

f) What did you learn from the plots above that you did not learn from the `summary()`? Explain.

The plot provides me with the information of central tendency of the distribution which I did not learn from the summary. While the summary does not let me know what the height of most of the babies is. In this bar chart, we can see that most of the babies's height is between 60 and 65.

g) Fill in the table below. Show the code you used in the R chunk below as well.

Variable	mean (sd) or n(/%)
Mother's Age	mean 27.25527 sd(5.781405)
Parity	mean 0.2548544 sd(0.4359557)
Gestation	mean 279.3385 sd(16.02769)
Birth weight (oz)	mean 119.5769 sd(18.23645)
Mother's weight (lbs)	mean 128.6258 sd(20.97186)

Variable	mean (sd) or n(/%)
Smoke status	no_smoke 742 (60.03%) smoke 484 (39.15%)

```
mean(babies$age, na.rm = TRUE)
```

```
## [1] 27.25527
```

```
sd(babies$age, na.rm = TRUE)
```

```
## [1] 5.781405
```

```
mean(babies$parity)
```

```
## [1] 0.2548544
```

```
sd(babies$parity)
```

```
## [1] 0.4359557
```

```
mean(babies$gestation, na.rm = TRUE)
```

```
## [1] 279.3385
```

```
sd(babies$gestation, na.rm = TRUE)
```

```
## [1] 16.02769
```

```
mean(babies$bwt)
```

```
## [1] 119.5769
```

```
sd(babies$bwt)
```

```
## [1] 18.23645
```

```
mean(babies$weight, na.rm = TRUE)
```

```
## [1] 128.6258
```

```
sd(babies$weight, na.rm = TRUE)
```

```
## [1] 20.97186
```

```
table(babies$smoke_f)
```

```
##
##  no yes
## 742 484
```

```
library(dplyr)
babies %>%
    group_by(smoke_f) %>%
    summarise(percent = 100 * n() / nrow(babies))
```

```
## # A tibble: 3 x 2
##   smoke_f percent
##   <fct>     <dbl>
## 1 no       60.0
## 2 yes      39.2
## 3 <NA>      0.809
```