

STAT 630: Homework 1

Brandon Keck

Due: September 6th, 2024 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? ~The Census Bureau reduced respondent burden by streamlining their surveys. They have accomplished this by avoiding asking redundant questions that could be linked to responses with administrative data. They have also included incentives and follow-up strategies to encourage participation without respondents feeling overwhelmed.
 - d) Describe the sampling process in a few sentences *in your own words*.
~The American Housing Survey works with the U.S. Census Bureau and uses their Master File. This master file contains a comprehensive list of housing units. A sample which represents approximately 100,000 housing units is then selected. This sample over samples larger metropolitan areas such as L.A. but it also captures smaller areas such as Richmond Virginia to ensure national and regional representatives.
 - e) What would you like to know about? Write a research question that could be answered with the American Housing Survey.
I would like to know what the housing market trend has been over the years. I would like to see if we could predict the future cost of homes will be in the next 10, 20, and 30 years and to see how many homes would need to be built in order to keep housing more affordable within the US.
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 from our book, load it, i.e., open the app!

```
library(openintro) # Load the openintro package
require(tidyverse) # Load the tidyverse suite of packages
```

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

```
data(babies) # Load the data

# Uncomment the line below to view the help file.
# ?babies # Make sure to comment back before knitting
```

View a summary of the dataset. Note: in all future assignments do NOT print the output from `glimpse()` or `summary()`. If you are going to provide summary statistics, they should appear in a neatly organized table.

```
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, ~
## $ 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, ~
```

```
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
```

- What does each row in the data frame represent, i.e., what is the observational unit?
~The observational unit of interest is each individual baby. This includes attributes that both relate to the baby and their mother. Each row of data contains information specific to each individual baby including; birth weight, gestation time, height and weight of the mother etc.
- How many participants were in the study? ~1,236 participants were in the study.
- All variables are coded as integers. Which variables should be recoded as *factors*? Recode these variables in the code chunk below. (Optional: do this with a single line of code.) ~The two variables that should be coded as factors are parity, and smoke. These are both categorical variables with two levels of 0 and 1.

```
babies$parity <- factor(babies$parity, levels = c(0, 1), labels = c("No", "Yes"))
babies$smoke <- factor(babies$smoke, levels = c(0, 1), labels = c("Non-Smoker", "Smoker"))
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    <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, ~
## $ 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     <fct> Non-Smoker, Non-Smoker, Smoker, Non-Smoker, Smoker, Non-Smok~
```

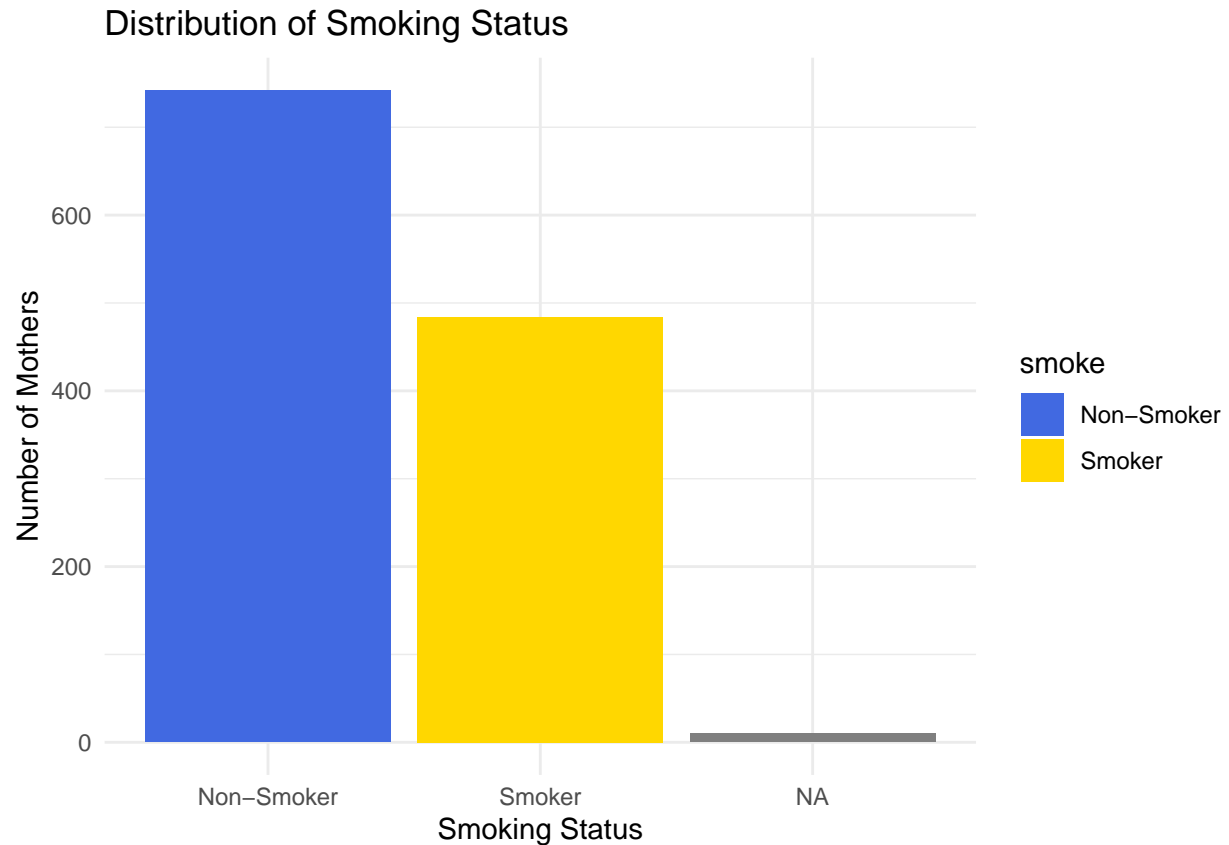
```
summary(babies) # View a summary of each column (variable)
```

```
##      case      bwt      gestation      parity      age
## Min.   : 1.0    Min.   : 55.0    Min.   :148.0    No :921    Min.   :15.00
## 1st Qu.:309.8   1st Qu.:108.8   1st Qu.:272.0   Yes:315   1st Qu.:23.00
## Median :618.5   Median :120.0   Median :280.0           Median :26.00
## Mean   :618.5   Mean   :119.6   Mean   :279.3           Mean   :27.26
## 3rd Qu.:927.2   3rd Qu.:131.0   3rd Qu.:288.0           3rd Qu.:31.00
## Max.   :1236.0   Max.   :176.0   Max.   :353.0           Max.   :45.00
##                                     NA's   :13           NA's   :2
##      height      weight      smoke
## Min.   :53.00    Min.   : 87.0    Non-Smoker:742
## 1st Qu.:62.00    1st Qu.:114.8   Smoker    :484
## Median :64.00    Median :125.0   NA's      : 10
## Mean   :64.05    Mean   :128.6
## 3rd Qu.:66.00    3rd Qu.:139.0
## Max.   :72.00    Max.   :250.0
## NA's   :22      NA's   :36
```

- d) Create a plot using base R or the tidyverse to visualize one categorical variable of your choice in the code chunk below. *Make sure to add a title and relabel the x and y axes.*

```
library(ggplot2)

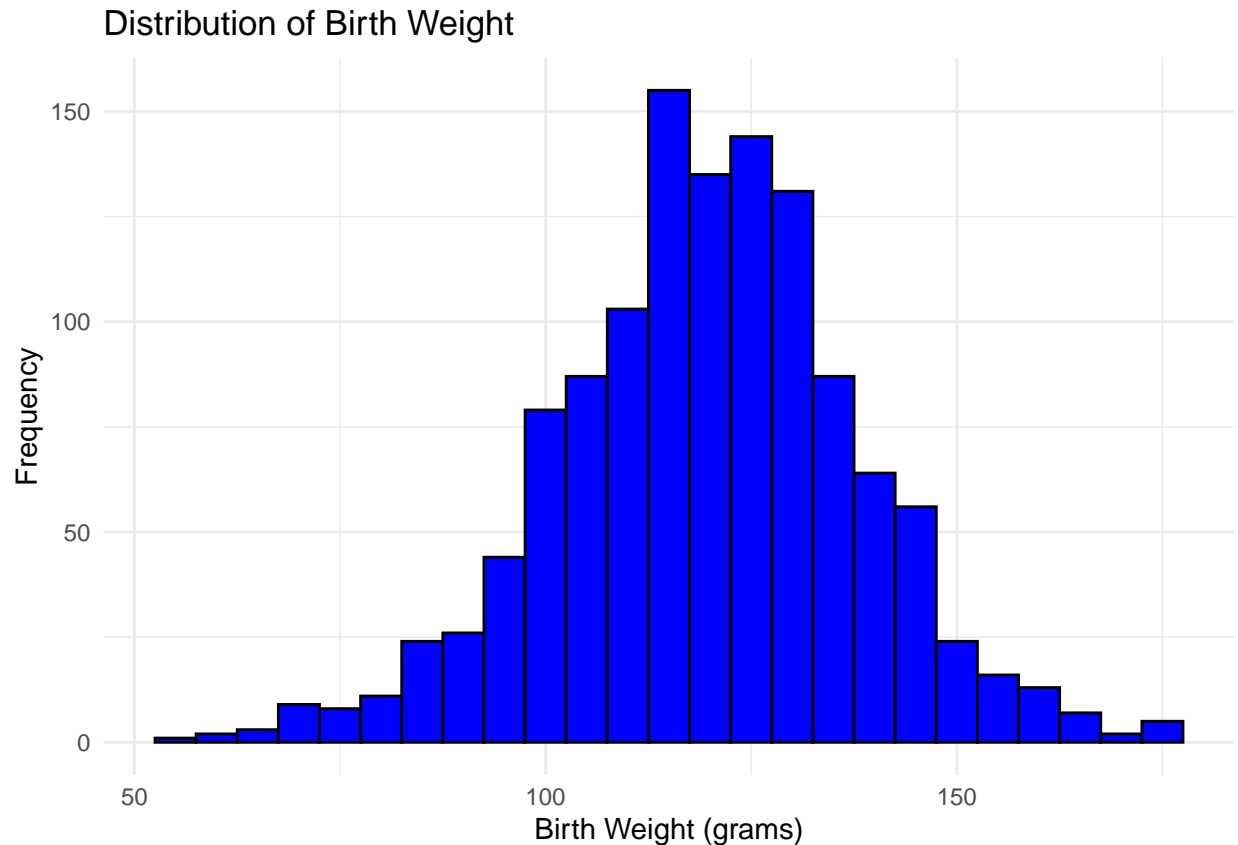
ggplot(babies, aes(x = smoke, fill = smoke)) +
  geom_bar() +
  scale_fill_manual(values = c("Non-Smoker" = "royalblue", "Smoker" = "gold")) +
  labs(title = "Distribution of Smoking Status",
       x = "Smoking Status",
       y = "Number of Mothers") +
  theme_minimal()
```



- e) Create a plot using base R or the tidyverse to visualize one quantitative variable of your choice in the code chunk below. *Make sure to add a title and relabel the x and y axes.*

```
library(ggplot2)

#histogram
ggplot(babies, aes(x = bwt)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribution of Birth Weight",
       x = "Birth Weight (grams)",
       y = "Frequency") +
  theme_minimal()
```



f) What did you learn from the plot of the quantitative variable in part (e) above that you did not learn from the summary statistics? This histogram of the distribution of birth weight shows a bell-shaped curve which indicates a normal distribution.

g) Manually fill in the table below (Optional: use code to generate your own table). Round to 2 decimal places. Show any code you used in the R chunk below as well.

```
mean_age <- mean(babies$age, na.rm = TRUE)
sd_age <- sd(babies$age, na.rm = TRUE)

# Parity table
props1 <- prop.table(table(babies$parity))
addmargins(table(babies$parity))
```

```
##
##   No  Yes  Sum
##  921  315 1236
```

```
mean_ges <- mean(babies$gestation, na.rm = TRUE)
sd_ges <- sd(babies$gestation, na.rm = TRUE)

mean_bwt <- mean(babies$bwt)
sd_bwt <- sd(babies$bwt)

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

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

```
# Smoking table
```

```
props2 <- prop.table(table(babies$smoke))
```

```
addmargins(table(babies$smoke))
```

```
##
```

```
## Non-Smoker      Smoker      Sum
```

```
##           742           484      1226
```

Variable	mean (sd) or n(%)
Mother's Age	27.26 (5.78)
Parity	
- 0	921 (74.51%)
- 1	315 (25.49%)
Gestation	279.34 (16.03)
Birth weight (oz)	119.58 (18.24)
Mother's weight (lbs)	128.63 (20.97)
Smoker Status	
- Non-Smoker	742 (60.52%)
- Smoker	484 (39.48%)

In the space provided, please write your answers to the following:

1) **Was there anything you found difficult with this homework? What topics (if any) do you feel you still**

The part that I found to be particularly difficult with this homework was the re-coding of the parity and smoker status variables to factor variables. It took me a while to understand that running the same code multiple times would make R confused. Sometimes I would run the code and the factors would become 0s and 1s respectfully. Then if I ran it again, it would become NA's which is not what I wanted them to be. In fact, it was going to Dr. Moore's office hours that helped alleviate that issue. The other part that I had difficulty with was manually filling in the table. I had initially attempted to create my own table but that was too much for me to take on.

2) **Give yourself a rating for this assignment using the EMRN rubric.** I feel that I earned an E for Excellent on this assignment. I spent many hours combing over my code and the questions to ensure the accuracy of my code and the understanding of each question. I also went to office hours to ensure my understanding of the topics that I was unsure of. Dr. Moore had even said that my code and theirs was identical. I feel that I have completed the assignment with all questions being answered to their entirety and with great accuracy. The only thing I believe that should make this an M would be because I didn't use code to generate my own table.