# STAT 630: Homework 1

David Teng

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?

Sampling Unit in the American Housing Survey: The American Housing Survey uses housing units as the sampling unit. It surveys the same housing unit every two years, collecting data on the physical characteristics and demographics of residents, even if they change.

b) What does HUD stand for?

HUD stands for the Department of Housing and Urban Development.

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

The Census Bureau reduces respondent burden by utilizing administrative data, such as public property records, to pre-fill information (e.g., the year a house was built), avoiding redundant or complex questions for respondents.

d) Describe the sampling process in a few sentences *in your own words*.

Sampling Process: The American Housing Survey picks addresses from a huge database called the Census Bureau's Master Address File. It makes sure the selection represents the whole country. About 100,000 housing units are chosen, and for bigger cities, they pick extra units to get better data. Smaller areas also get sampled every few years so they can track housing changes all over the U.S. This helps them create a complete picture of housing in different regions

e) What would you like to know about? Write a research question that could be answered with the American Housing Survey.

How have changes in the housing market, especially post-pandemic, affected the quality of life for people in large metropolitan areas? Are more people living together in shared homes, and how has this impacted their housing conditions and costs?

This question could help explore trends like the rise of unrelated families living together and how housing affordability and availability have evolved, which are key focuses of the American Housing Survey. It could reveal valuable insights for policymakers to improve housing conditions.

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.*

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

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

```r
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.

```r
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.

```r
glimpse(babies) # Glimpse the dataset
summary(babies) # View a summary of each column (variable)
```

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

Each row in the babies dataframe represents a single birth, where variables are recorded for each birth event.

b) How many participants were in the study?

```r
data(babies)
nrow(babies)
```

```
## [1] 1236
```

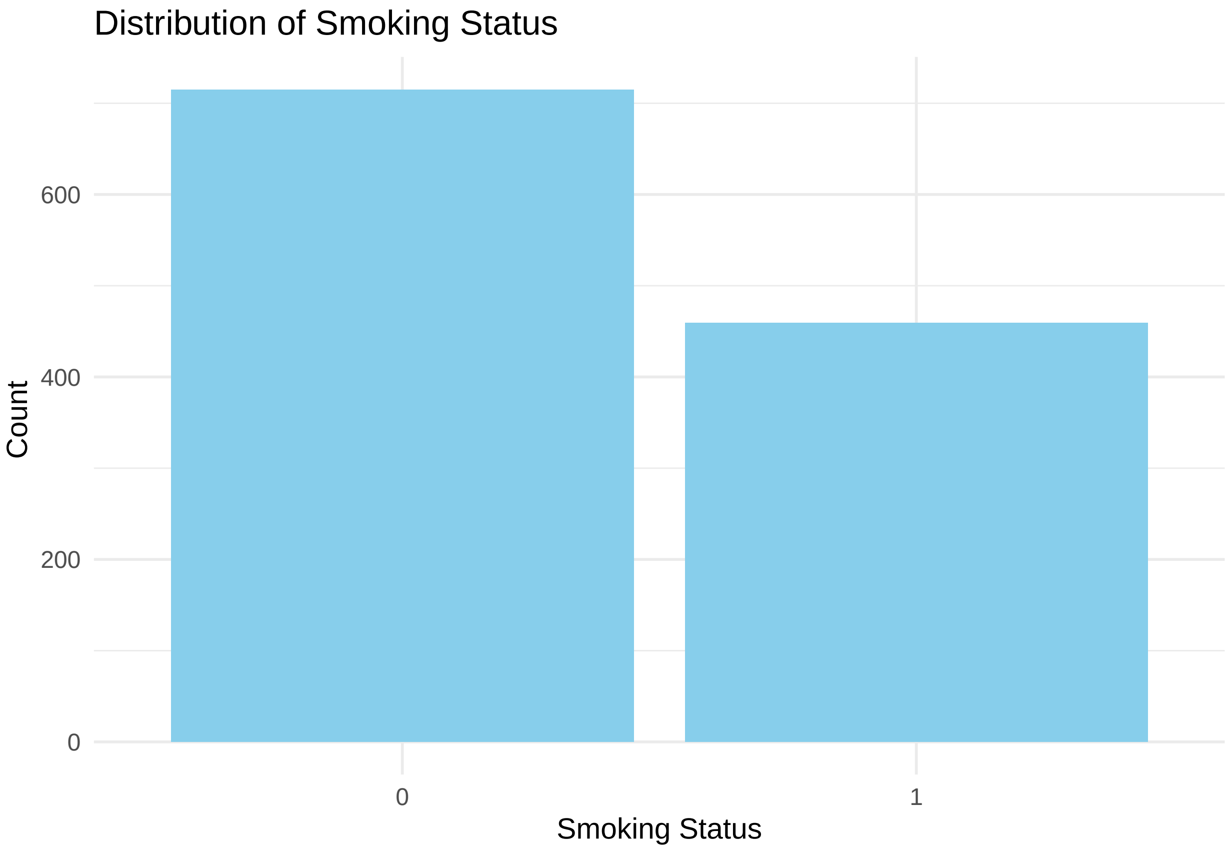
c) All variables are coded as integers. Which variables should be recoded as *factors*? Recode these variables in the code chunk below. (Optional: do this with a single line of code.)

```r
babies <- babies %>%
  mutate(parity = as.factor(parity),
         smoke = as.factor(smoke))
babies
```

```
## # A tibble: 1,236 x 8
##     case   bwt gestation parity   age height weight smoke
##    <int> <int>     <int> <fct>  <int>  <int>  <int> <fct>
##  1     1   120       284 0         27     62    100 0
##  2     2   113       282 0         33     64    135 0
##  3     3   128       279 0         28     64    115 1
##  4     4   123        NA 0         36     69    190 0
##  5     5   108       282 0         23     67    125 1
##  6     6   136       286 0         25     62     93 0
##  7     7   138       244 0         33     62    178 0
##  8     8   132       245 0         23     65    140 0
##  9     9   120       289 0         25     62    125 0
## 10    10   143       299 0         30     66    136 1
## # i 1,226 more rows
```

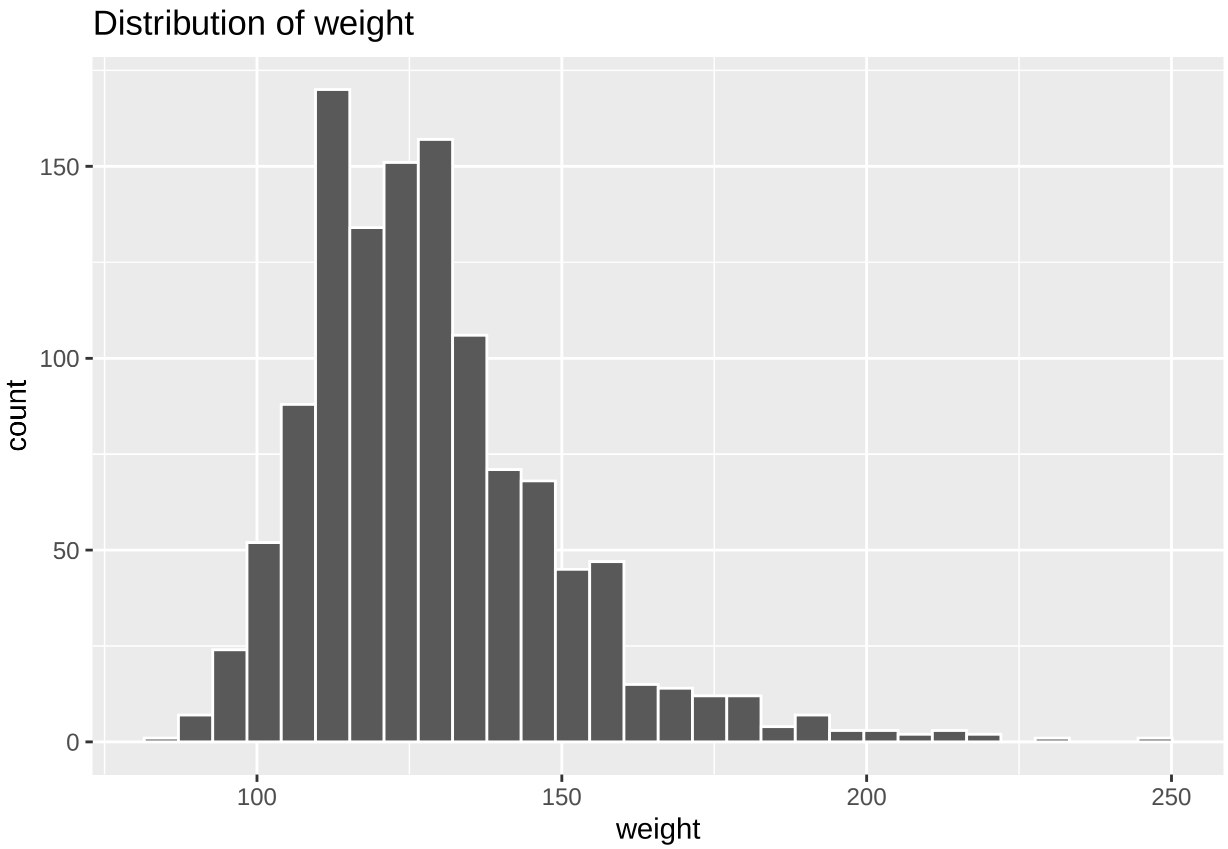
d) Create a plot using base R or the tidyverse to visualize one categorical variable of your choice in the code chunk below. *Make sure to add a title and relabel the x and y axes.*

```r
na.omit(babies) %>%
ggplot(aes(x = smoke)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of Smoking Status",
       x = "Smoking Status",
       y = "Count ") +
  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.*

```
babies %>%
  ggplot(aes(x = weight)) +
    geom_histogram(bin = 10, color= "white")+
labs(title = "Distribution of weight",
       x = "weight",
       y = "count")
```



Distribution of weight

f) What did you learn from the plot of the quantitative variable in part (e) above that you did not learn from the `summary()`? Explain.

```
summary(babies$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    87.0   114.8   125.0   128.6   139.0   250.0      36 
```

Plot: The plot may reveal patterns, trends, outliers, skewness, or distribution shape that are not immediately evident from summary statistics alone.

Summary: The summary() function provides basic statistics such as the mean, median, quartiles, and extremes (min, max).

Hence, compared with “summary”, ’plot” provides a visual sense of data distribution, which can be crucial for understanding the underlying data structure.

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.

| Variable | mean (sd) or n(%) |
|---|---|
| Mother’s Age | |
| Parity | |
| Gestation | |
| Birth weight (oz) | |
| Mother’s weight (lbs) | |
| Smoke status | |

```
library(gtsummary)
library(dplyr)

data(babies)

summary_table <-
  babies %>%
  select(age,
         parity,
         gestation,
         bwt,
         weight,
         smoke) %>%
  tbl_summary(
    statistic = list(all_continuous() ~ "{mean} ({sd})", all_categorical() ~ "{n} ({p}%)"))

summary_table
```

| **Characteristic** | **N = 1,236**[1] |
|---|---|
| age | 27 (6) |
| Unknown | 2 |
| parity | 315 (25%) |
| gestation | 279 (16) |
| Unknown | 13 |
| bwt | 120 (18) |
| weight | 129 (21) |
| Unknown | 36 |
| smoke | 484 (39%) |
| Unknown | 10 |

[1] Mean (SD); n (%)

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

For part (g), I found this task quite challenging, particularly because I needed to generate a table with the calculated statistics. However, through my efforts and perseverance, I was able to complete it. This experience not only enhanced my knowledge but also gave me a profound sense of achievement.

2) Give yourself a rating for this assignment using the EMRN rubric.

E - Excellent