R Notebook

Conceptual Questions

1. (4 points) For each of the following, determine whether the response variable numerical or categorical. If the response variable is categorical, is it binary? If it is not binary, list possible categories for the response variable.

(a) In a survey, college students were asked how many hours per week they spend on the internet.

Numerical

(b) In a survey, college students were asked, "What percentage of the time that you spend on the internet is not for course work?"

Numerical

(c) In a survey, college students were asked, "What is your primary mode of transportation when traveling between campus and home?"

Categorical non-binary: Car, Walk, Cycle, Motorcycle, bus, Etc.

(d) In a survey, college students were asked whether or not they live on campus.

Categorical binary: Yes/no

(e) In a survey, college students were asked how many of their meals they prepare at home per week.

Numerical

(f) In a survey, college students were asked, "Which of the five main food groups constitutes the majority of your diet?"

Categorical non-binary: Fruits, Vegatables, Grains, Protein, Dairy

(g) In a survey, smart phone users were asked whether or not they have used a web-based taxi service like Uber or Lyft.

Categorical binary: yes/no

(h) In a survey, smart phone users were asked how many times they used a web-based taxi service in the past three months.

Numerical

2. (3 points) On a multiple-choice exam, each of the 20 questions has 2 possible answers and only one correct response. Suppose, for each question, one student selects his responses completely at random.

(a) Let Xi, i = 1, . . . , 20 represent whether the student answered the ith question correctly, and let Y represent the total number of answers the student gets correct. What is the distribution of Xi? What is the distribution of Y ?

Xi will have a bernoulli distribution and Y will have a binomial distribution.

(b) What is the probability that the student passes the test (i.e., scores at least 70%)? (You will learn more about this distribution in the next module, but you should be able to calculate this in R using the pbinom function.)

```
pbinom(14,20,0.50, lower.tail = F)
```

```
## [1] 0.02069473
```

Probability that student scores higher than 70% is roughly 2.1%.

3. (2 points) Do you prefer taking courses online, on campus, or a hybrid of the two? Suppose these preferences occur with probabilities (p1, p2, p3). For N = 3 independent subjects, let the observed frequencies be (n1, n2, n3). That is, we observe n1 subjects who prefer taking courses online, etc.

(a) Explain how you can determine n3 from knowing n1 and n2.

If there are N total participants in the study split between n1, n2, and n3, then the equation N = n1+n2+n3 would be true. Therefor, if n1 and n2 are known, with N given above, it's a simple matter of re-arranging the formula and calculating n3: n3 = N - (n2 + n1)

(b) List all the possible observations (n1, n2, n3), with n = 3.

i	n1n2n3
1	0 0 3
2	0 1 2
3	1 1 1
4	1 0 2
5	3 0 0
6	2 1 0
7	2 0 1
8	0 3 0
9	0 2 1
10	1 2 0

4. (1 point) Consider the data in ex2117 of the Sleuth3 package.

(a) In what format are these data (case, tabular, frequency, other)? Please explain.

```
library('Sleuth3')
```

```
## Warning: package 'Sleuth3' was built under R version 4.0.3
```

```
df <- ex2117
df
```

```
##     Group           Time Number PctBoys
## 1 Control           none  20337    51.2
## 2 Exposed       13-16mo     71    52.1
## 3 Exposed        12-7mo    789    49.6
## 4 Exposed         0-6mo   1922    48.9
## 5 Exposed 1st trimester    290    46.0
```

These data are formatted in a frequency data format with an additional column for percentages.

(b) Create a new file for the data, called HW1data. In that file, create a NumBoys column that represents the number of births that were boys (be sure to round to the nearest integer).

```
HW1data <- df

HW1data['NumBoys'] <- round(HW1data['PctBoys'] * HW1data['Number'] / 100,0)


write.csv(HW1data,"HW1Data.csv")
```

(c) Use an appropriate R command or commands to change the format to something else (you can pick what).

```
# CASE FORMAT
head(as.data.frame(table(HW1data)))
```

```
##     Group    Time Number PctBoys NumBoys Freq
## 1 Control   0-6mo     71      46      37    0
## 2 Exposed   0-6mo     71      46      37    0
## 3 Control  12-7mo     71      46      37    0
## 4 Exposed  12-7mo     71      46      37    0
## 5 Control 13-16mo     71      46      37    0
## 6 Exposed 13-16mo     71      46      37    0
```