Assignment/Homework 1 - Tam Nguyen

1.

Identify whether the following variables are numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the categories have a natural order (ordinal) or not (nominal).

a. Number of sexual partners in a year

numerical, discrete

b. Petal area of rose flowers

numerical, continuous

c. Heart beats per minute of a Tour de France cyclist, averaged over the duration of the race

numerical, discrete

d. Birth weight

numerical, continuous

e. Stage of fruit ripeness (e.g., underripe, ripe, or overripe)

categorical, nominal

f. Angle of flower orientation relative to position of the sun

numerical , continuous

g. Tree species

categorical, nominal

h. Year of birth

numerical, discrete

i. Gender

categorical, nominal

2.

Not all telephone polls carried out to estimate voter or consumer preferences make calls to cell phones. One reason is that in the USA automated calls (“robocalls”) to cell phones are not permitted, and interviews conducted by humans are more costly.

a. How might the strategy of leaving out cell phones affect the goal of obtaining a random sample of voters or consumers?

this might lead to bias. They are members that are difficult to collect, but leaving them out will lead to an under-representaion of the population

b. Which criterion of random sampling is most likely to be violated by the problems you identified in part (a): equal chance of being selected, or the independence of the selection of individuals?

equal chance of being selected - these difficult to collect individuals might have characteristics differed from the rest of the population, so they can be quite important.

c. Which attribute of estimated consumer preference is most affected by the problem you identified in (a): accuracy or precision?

this is related to accuracy - the result of the sample not properly taken.

3.

In each of the following examples, indicate which variable is the explanatory variable and which is the response variable.

a. The anticoagulant warfarin is often used as a pesticide against house mice, Mus musculus. Some populations of the house mouse have acquired a mutation in the vkorc1 gene from hybridizing with the Algerian mouse, M. spretus (Song et al. 2011). In the Algerian mice, this gene confers resistance to warfarin. In a hypothetical follow-up study, researchers collected a sample of house mice to determine whether presence of the vkorc1 mutation is associated with warfarin resistance in house mice as well. They fed warfarin to all the mice in a sample and compared survival between the individuals possessing the mutation and those not possessing the mutation.

Response variable: survival rate

Explanatory variable: the vkorc1 mutation

b. Cooley et al. (2009) randomly assigned either of two treatments, naturopathic care (diet counseling, breathing techniques, vitamins and a herbal medicine) or standardized psychotherapy (psychotherapy with breathing techniques and a placebo added), to 81 individuals having moderate to severe anxiety. Anxiety scores decreased an average of 57% in the naturopathic group and 31% in the psychotherapy group.

Explanatory variable: naturopathic care, standardized psychotherapy

Response variable: anxiety level

c. Individuals highly sensitive to rewards tend to experience more food cravings and are more likely to be overweight or develop eating disorders than other people. Beaver et al. (2006) recruited 14 healthy volunteers and scored their reward sensitivity using a questionnaire (they were asked to answer “yes” or “no” to questions like: “I’m always willing to try something new if I think it will be fun”). The subjects were then presented with images of appetizing foods (e.g., chocolate cake, pizza) while activity of their fronto–striatal–amygdala–midbrain was measured using functional MRI. Reward sensitivity of subjects was found to correlate with brain activity in response to the images.

Explanatory variable: activity of participants’ fronto–striatal–amygdala–midbrain

Response variable: reward sensitivity

d. Endostatin is a naturally occurring protein in humans and mice that inhibits the growth of blood vessels. O’Reilly et al. (1997) investigated its effects on growth of cancer tumors, whose growth and spread requires blood vessel proliferation. Mice having lung carcinoma tumors were randomly divided into groups that were treated with doses of either 0, 2.5, 10, and 20 mg/kg of endostatin injected once daily. They found that higher doses of endostatin led to inhibition of tumor growth.

Explanatory variable: doses of edostatin

Response variable: tumor growth

4.

For each of the studies presented in problem 3, indicate whether the study is an experimental or observational study.

1. house mice study: experimental study

2. anxiety study: experimental study

3. food craving study: observational study

4. tumor growth study: experimental study

5.

The Cambridge Study in Delinquent Development was undertaken in north London (UK) to investigate the links between criminal behavior in young men and the socioeconomic factors of their upbringing (Farrington 1994). A cohort of 395 boys was followed for about 20 years, starting at the age of eight or nine. All of the boys attended six schools located near the research office. The following table shows the total number of criminal convictions by the boys between the start and end of the study.

a. What type of table is this?

frequency table

b. How many variables are presented in this table?

there is only one numerical variable: number of convictions

c. How many boys had exactly two convictions by the end of the study?

21

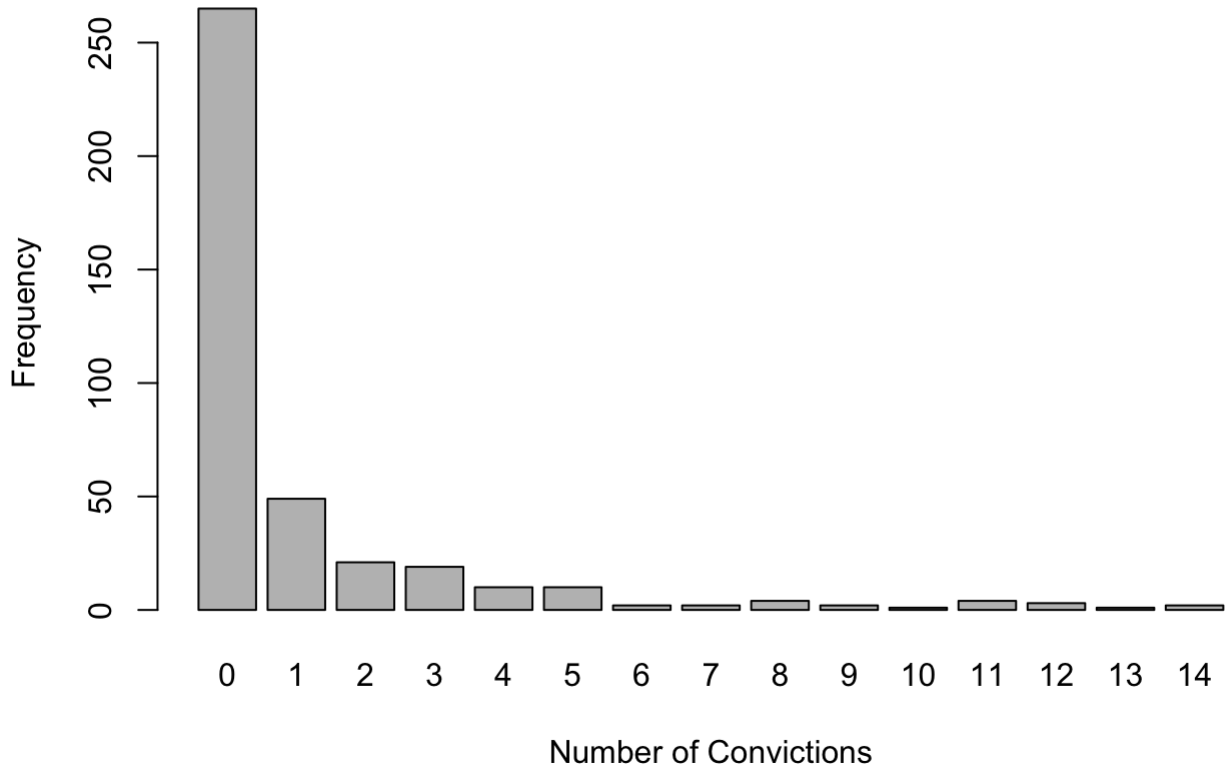
d. What fraction of boys had no convictions?

0.67 or 67% of boys have no conviction

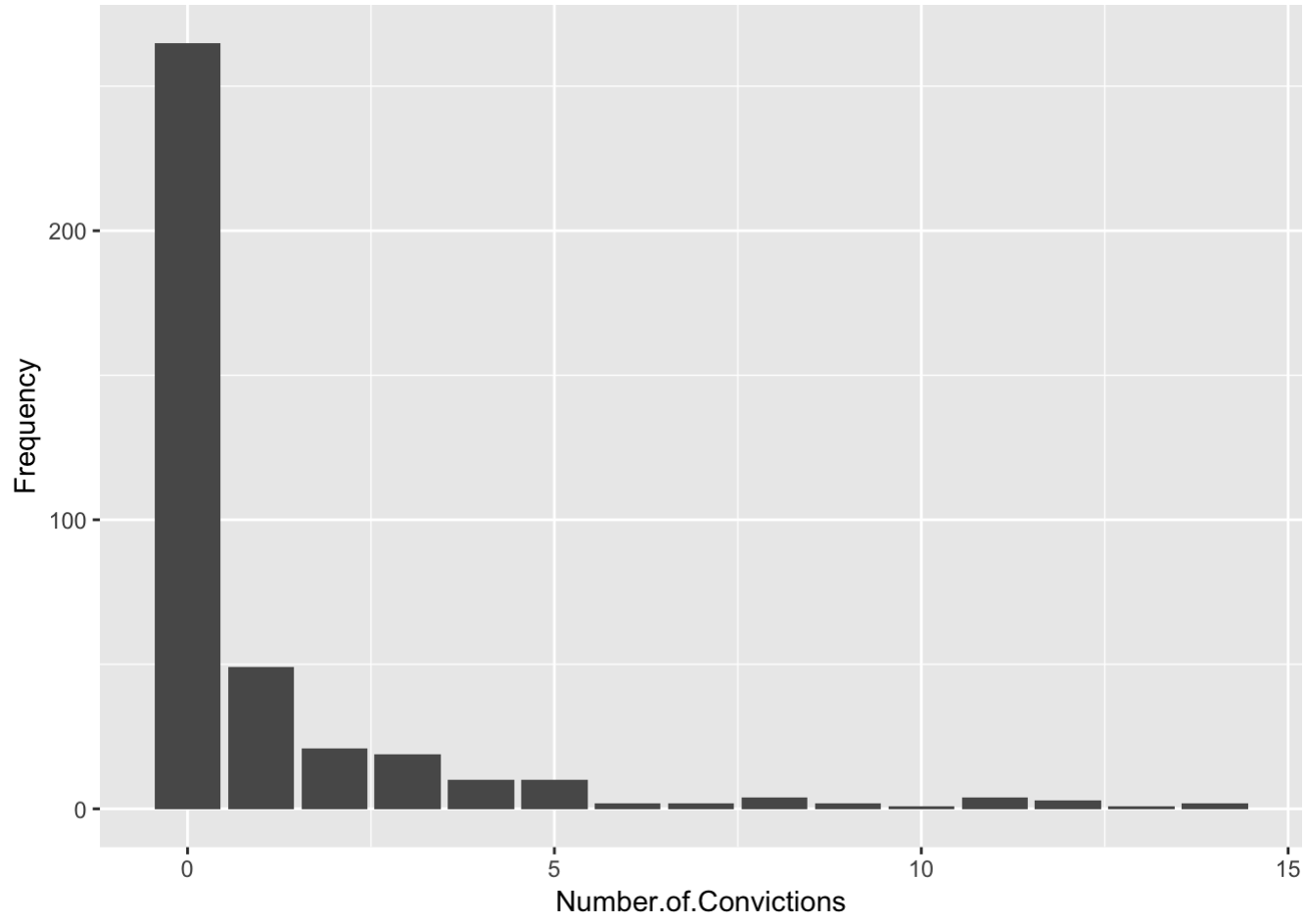
e. Display the frequency distribution in a graph. Which type of graph is most appropriate? Why?

```
graph1 <- data.frame("Number of Convictions" = 0:14, "Frequency" = c(265, 49, 21, 19,
 10, 10, 2, 2, 4, 2, 1, 4, 3, 1, 2))

barplot(graph1$Frequency,
        axisnames = TRUE,
        names.arg = graph1$Number.of.Convictions,
        xlab = "Number of Convictions",
        ylab = "Frequency")
```




```
ggplot(graph1, aes(Number.of.Convictions, Frequency)) +
  geom_col()
```



bar graph is the most appropiate since the number of convictions is a numeric and dicreet variable and it helps to visualise the frequency of each number of convictions.

f. Describe the shape of the frequency distribution. Is it skewed or is it symmetric? Is it unimodal or bimodal? Where is the mode in number of criminal convictions? Are there outliers in the number of convictions?

the shape is skewed and it is unimodal. The mode is 0 in terms of the number of criminal convictions. 0 conviction is also an outlier in the number of conviction because of it large frequency compared to the rest.

g. Does the sample of boys used in this study represent a random sample of British boys? Why or why not?

This sample does not represent a random sample because it is stated that all the boys are recruited in the study are all located near the research office. Boys in other areas are not recruited. So this lead to bias and affects the accuracy of the sample.

6.

The following graph was drawn using a very popular spreadsheet program in an attempt to show the frequencies of observations in four hypothetical groups. Before reading further, estimate by eye the frequencies in each of the four groups.

a. Identify two features of this graph that cause it to violate the principle, “Make patterns in the data easy to see.”

1. The angled perspective of the 3D graph makes it difficult to judge the bar height by eye. It makes the pattern harder to see.

2. The units in vertical axis cram too many numbers. It adds a number in every 5 points. Better to show a number every 10 points, like 0, 10, 20, 30, 40 instead of 0, 5, 10…

b. Identify at least two other features of the graph that make it difficult to interpret.

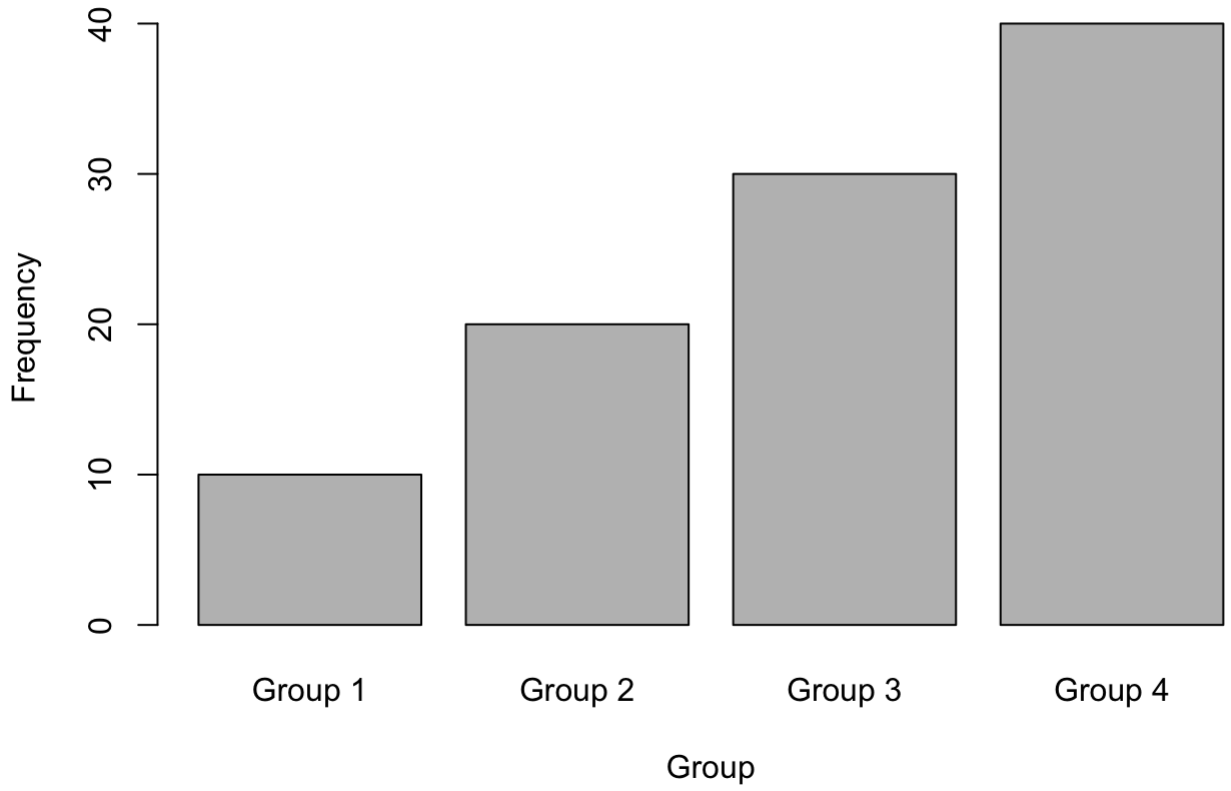
1. Graphical elements are not clearly labeled. The vertical axis is not labeled.

2. the colors, shapes and shadows of this graph make it overcomplicated while not communicating any further information. Colors in this graph are also not differ in intensity and shapes, making it hard for readers to distinguish between groups in the graph. The shape of this graph is shaped like cones, making it harder to know where is the exact point of frequency to line up with the vertical axis.

c. The actual frequencies are 10, 20, 30, and 40. Draw a graph that overcomes the problems identified above.

```
graph2 <- data.frame("Group" = c("Group 1", "Group 2", "Group 3", "Group 4"), "Freque
ncy" = c(10, 20, 30, 40))


barplot(graph2$Frequency,
        names.arg = graph2$Group,
        xlab = "Group",
        ylab = "Frequency")
```




```
ggplot(graph2, aes(Group, Frequency)) +
  geom_col()
```