

1. Suppose you have a data frame (or data table) called `houses`,

- (a) (3 pts) Suppose you run the commands below to obtain the results that follow them. What can you know about the sample and data contained in the data table? [Do not merely tell me that you know the numbers of rows and columns.]

```
nrow(houses)
```

```
[1] 8
```

There are 8 cases, and 15 variables measured

```
ncol(houses)
```

```
[1] 15
```

on those cases.

- (b) (4 pts) Given that the name of the data frame reveals what cases were measured, write the names of two possible variables, one that is categorical, one that is quantitative, which might appear in `houses`.

Variable Q (the quantitative one): square footage, price, number of bedrooms

Variable C (the categorical one): color, garage? (yes/no), city (location)

- (c) (3 pts) Name an appropriate graph type (or other "display") which is useful for viewing the distribution of your Variable C.

A frequency table or a bar graph

- (d) (3 pts each) Suppose one of the questions for which the data in `houses` is being collected is, "What is the average value of Variable Q for houses in the city of Grand Rapids?" Here are some ways houses might be sampled. Indicate whether the sampling method described is biased (B) or unbiased (U). If biased, describe in what way?

- B Out of the multiple (26 or so) zip code regions in Grand Rapids, just one is selected at random, and the territory it designates is split up into distinct square-mile-sized blocks, with 4 houses selected at random from each block. Thus, if there were 11 packets in the selected zip code, you would have $4 \times 11 = 44$ houses in your sample, all coming from one zip code.

Some zip codes may be over-representing houses of a certain type or value. Expensive homes tend to cluster, and may tend to be built in one zip code over another.

- U Only the houses whose owner's have a last name beginning with the letter "L" are included in the sample.

A possible reason for opposing this answer is the argument that this letter is more prevalent to one ethnicity than to some others. Presumably, this ethnicity would also tend to own different sorts of homes than others do.

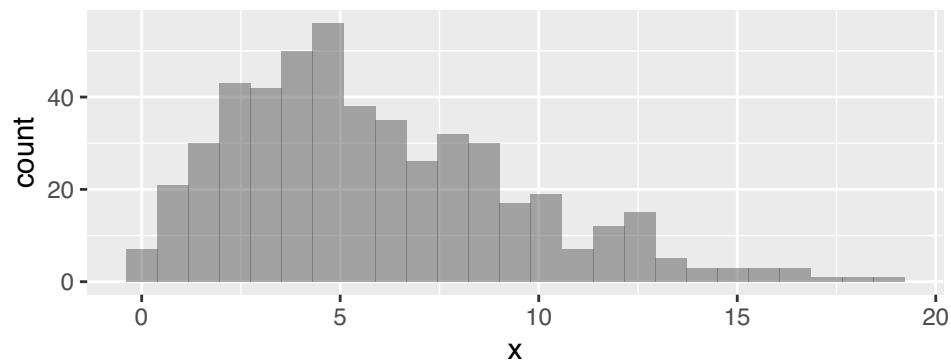
- B Only the houses whose owner's have signed-in (indicating they are present in the gallery) at a meeting of the City Council during the 2024 calendar year are included in the sample.

Very often it is only people with a complaint or vested interest who attend these meetings. The interest in civic affairs may suggest an association between attendance (yes/no) and the kind of house a person owns.

2. (2 pts each) Indicate if the study described is an experiment (E) or an Observational study (O).

- (a) O The subjects recruited for the study are asked whether they take a multivitamin with the understanding that they should continue their usual practice for two months, at which time they are asked questions about their general health over the two months. The goal is to see what link might exist between consuming a multivitamin and health outcomes.
- (b) E Dorm students are recruited, and some are given noise-cancelling headphones to wear while sleeping. The rest are not allowed any such device on their heads during sleeping hours. After two months, the average number of hours of sleep is compared to see what link might exist between the presence of such headphones and duration of sleep.

3. Look at the histogram provided.



(a) (4 pts) Describe the shape/form of the distribution.

The distribution is unimodal (blur it) and left-skewed.

(b) (3 pts) For this data, the mean is greater than ("less than", "equal to", "greater than") the median.