

1. Suppose we are scientists who study rabbits (bunny biologists), and we're interested in the burrowing behavior of the rabbits in a field. Specifically, we would like to know the *smallest distance* between any of the rabbit holes, among all pairs of holes. The holes are hidden, so we can't see them directly, but we can see the rabbits when they poke their heads out. By spraying the field with a carrot-based perfume, we entice some of the rabbits to poke their heads out of their holes, and we note the locations of those holes. Suppose these locations constitute a random sample without replacement of 30 of the hole locations, out of 100 total holes.

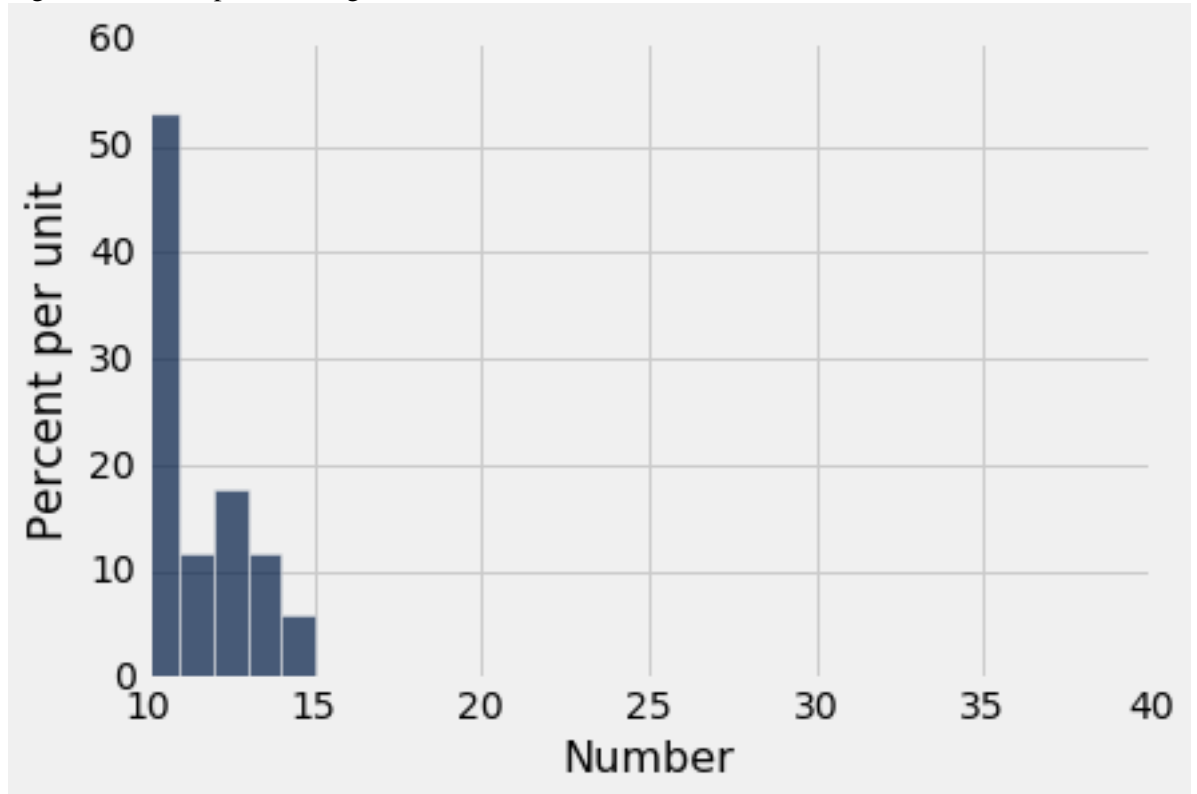
We record these locations in a table called `hole_location_sample`. It has two columns: `x`, the `x` coordinate of each hole, and `y`, the `y` coordinate of each hole. The distance between two holes can be calculated using their coordinates and the Pythagorean theorem.

- (a) Using only our sample, what would be an appropriate statistic to use as an estimate of the smallest distance between any two rabbit holes in the field? (For example, "the average distance between holes in the sample" would be a (poor) estimate.)
- (b) If you wanted to know how well your estimate worked, you could repeat our sampling process 1,000 times and make a histogram of the resulting 1,000 estimates. What would that histogram be called?
- (c) Suppose the smallest distance between any two holes in the field is 10 meters. Would you expect the histogram in (b) to have its mean at roughly 10 meters, or less than 10 meters, or more than 10 meters? Draw what you think it would look like.
- (d) As Patrick Star once said, "We're not cavemen! We have technology!" Suppose we have access to *all 100* of the hole locations in a table called `hole_locations`, and a function called `estimate_smallest_distance` that computes your proposed estimate on 1 sample. Write code that will generate the histogram in (b).

**Answers:**

- (a) A lot of correct answers here, but the one that makes the most sense is the minimum distance amount all of the rabbit wholes we see.
- (b) This would be an empirical histogram.

- (c) By doing this 1000 times, we know that a majority of our values will be around the area of 10 meters. The only problem is, however, is that we can never get too low, we can only get too big. We would expect a histogram like this:

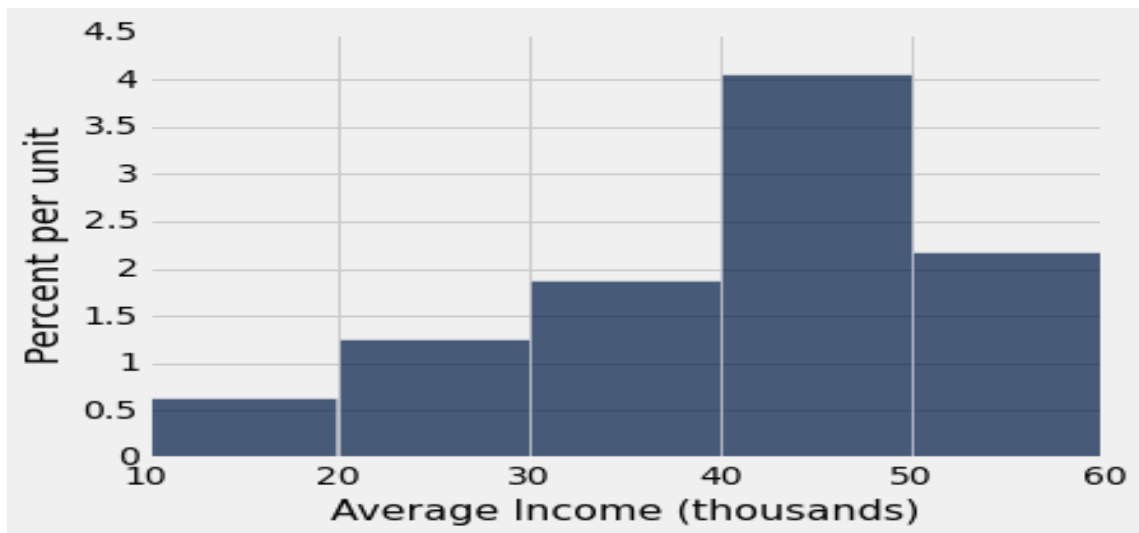


That is, a lot of values are 10, but there are some values greater than 10, none less though. Hence, our mean will end up being greater than 10.

- (d)
- ```
def estimate_smallest_difference :
    sample = hole_locations.sample(30)
    return np.max(sample.column(0)) -
           np.min(sample.column(0))
```

2. Below is a histogram of a the income distribution of a small city.

- (a) Which of the following seems closest to the mean? A. 35 B. **42** C. 51 D. 59
- (b) Which of the following seems closest to the standard deviation? A. 1 B. 5 C. **15** D. 30
- (c) Should we expect the mean to be the same, **less than**, or greater than the median?
- (d) Is this graph **skewed-left**, skewed-right, or symmetric?
3. The Data 8 staff decide that they would like to get a sense of whether or not tutoring is an effective way to raise the scores of students. They keep track of the students who do two on one tutoring and compare their midterm scores to those that do not participate in tutoring. They find that on average, students who participated in tutoring sessions did 25 percent better on the midterm than students who did not.



- Does this study have a treatment group and a control group? If so, name them.
- Is this an observational study, a randomized controlled experiment, or neither?
- The staff are convinced that the tutoring sections are causing the increase in midterm scores. Do you agree or disagree? Why or why not? If you agree, justify your answer. If not, give a specific reason as to why.

#### Answers:

- Yes, the treatment group is those who are receiving the mentoring and the control group is those who are not.
  - Observational Study, the staff is simply watching people and is not messing with the experience at all.
  - No, many confounding factors exist. One could be that students who are going to tutoring are more motivated to study and try and do well in the class, so they spend more hours outside of tutoring studying for the midterm. In general, causation can not be determined from an observational study.
4. Assume we are given a table called `scores` which has 10 columns , the first one containing the names of students, and the next nine containing their scores on quizzes 1-9 (in order). Design a function called `performance` which takes in a person's name and outputs a table with two columns, one for the quiz number and the other with a boolean value of either `true` or `false` indicating whether the student performed better than average on that quiz.  
Hint: Try using a `for` loop.

**Answers:**

```
def performance(name):
    personal = scores.where('name', name)
    t = Table.with_columns([
        'Quiz Number', [],
        'Better than Avg?', []
    ])
    for x in range(1,10):
        score = personal.column(x).item(0)
        avg = np.mean(scores.column(x))
        better = score > avg
        t.append(["Quiz " + str(x), better])
    return t
```

5. Let's say we have used the above function and assigned Marissa to the call `performance('Marissa')`. Using the table we now have, how can we use code to figure out how many tests Marissa did better than average on? There should be a relatively easy and concise solution.

**Answers:**

```
np.count_nonzero(marissa.column(1))
```