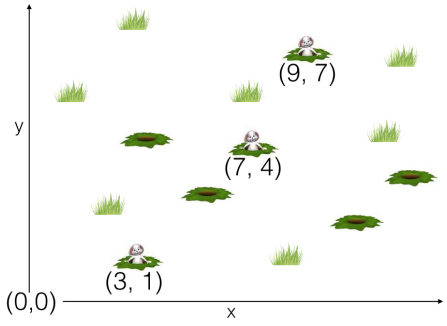
Discussion 8: Midterm Review

1. Suppose we are interested in the burrowing behavior of 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.



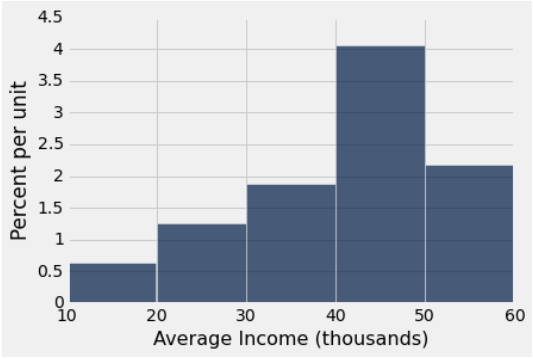
   (a) Using only our sample, what would be an appropriate statistic to use as an estimate of the smallest distance between any two rabbit holes in the field? (For example, "the average distance between holes in the sample" would be a (poor) estimate.)

   (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`. Suppose also that you have already written a function called `estimate_smallest_distance` that takes a table of 30 hole locations and returns the value of your proposed estimate computed on that sample. Write code that will generate the histogram described in (b).

2. Below is a histogram of 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) Is the mean the same, less than, or equal to the median, or can you not tell from this histogram?

(d) Is this histogram skewed-left, skewed-right, or symmetric?

3. The Data 8 staff decide 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 participate in two-on-one tutoring and compare their midterm scores to those who do not participate in tutoring. They find that on average, students who participate in tutoring sessions have midterm scores 25 percent higher than the midterm scores of students who do not.

(a) Does this study have a treatment group and a control group? If so, name them.

(b) Is this an observational study, a randomized controlled experiment, or neither?

(c) 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.

4. The table `scores` 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 as its argument a person's name and returns 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.

5. Using the above function, we write `Marissa = performance('Marissa')`. Write code to figure out how many tests Marissa did better than average on. There is a concise solution.