Week 7: Sampling and Hypothesis Testing

Data 8 Tutoring

# 1 Sampling

## Key Concepts

**Population vs. Sample**

In data science, we often want to be able to make a general statement about a **population** of individuals. Unfortunately, resource constraints generally prevent scientists from having access to data about entire populations of individuals. For that reason, we examine parts of the population called **samples**. Our goal is to infer some characteristics of the population, called **population parameters** from the study of our sample. In many cases, we are interested in estimating these parameters using **sample statistics**, or quantities that we measure from a sample of the population.

For example, you may be interested in knowing the percentage of all eligible voters who are registered to vote for the upcoming election. Since asking everyone in the U.S. if they have registered to vote is clearly infeasible, we will have to take a sample.

Sometimes, we will want to sample from a pre-existing table. To do so, we can use the following table method:

tbl.sample(sample\_size)

In other cases, we may have an array we need to sample from. In this case, we can use the following function:

np.random.choice(array, sample\_size)

## Practice Problems

**1.1** Let’s use the example of rolling a fair die. Remember: rolling a die is always sampling “with replacement”.

1. What is the probability that you will roll a 5? Is this an empirical or a theoretical probability? Is there a relationship between the two?



1. Complete the function roll\_die, which takes in no arguments and uses the dice table to the right to roll a dice a single time and returns the value that is randomly picked.

def roll\_die():

1. Simulate rolling a die 10 times and store the results in an array called simulated\_rolls.

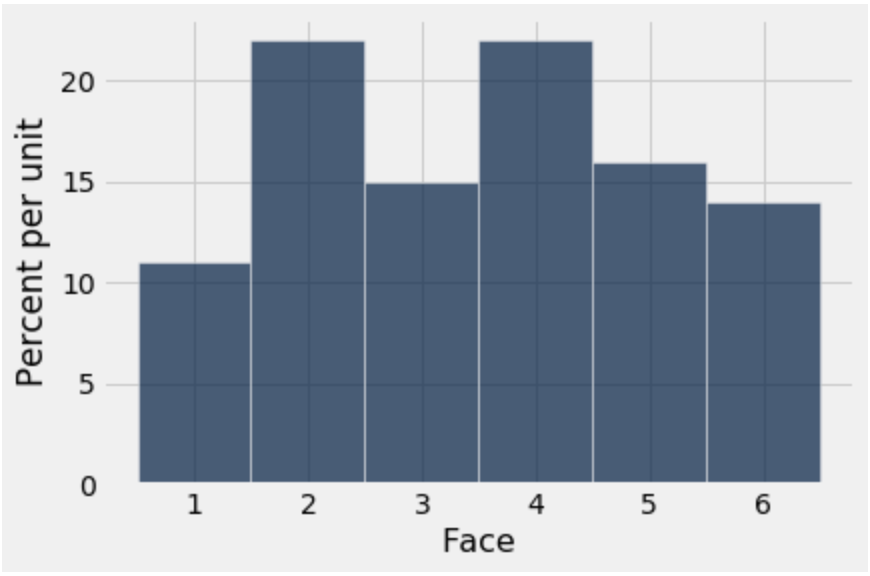
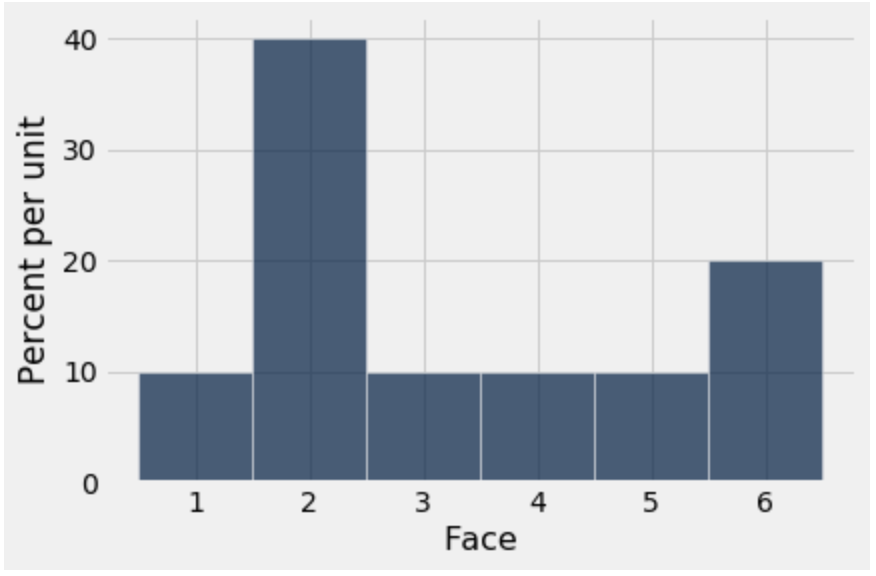
simulated\_rolls = make\_array()

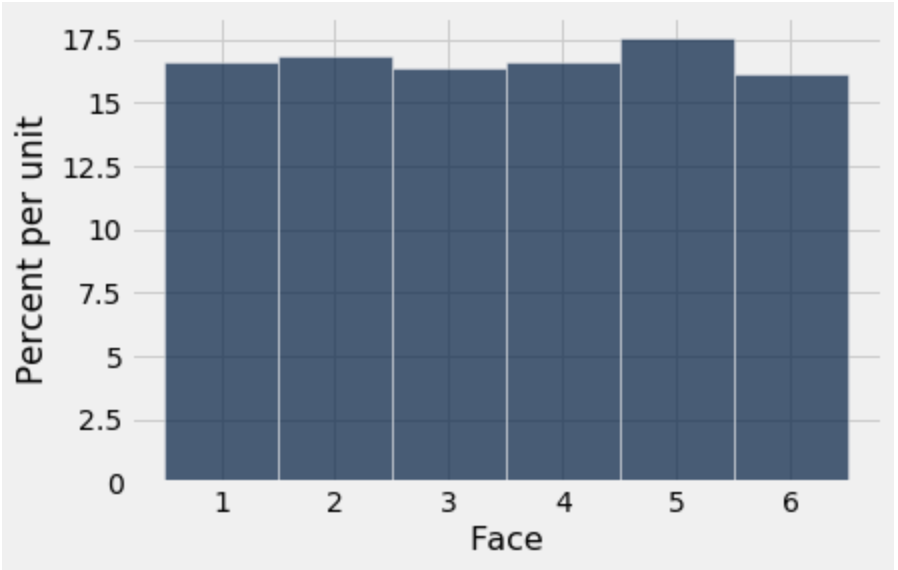
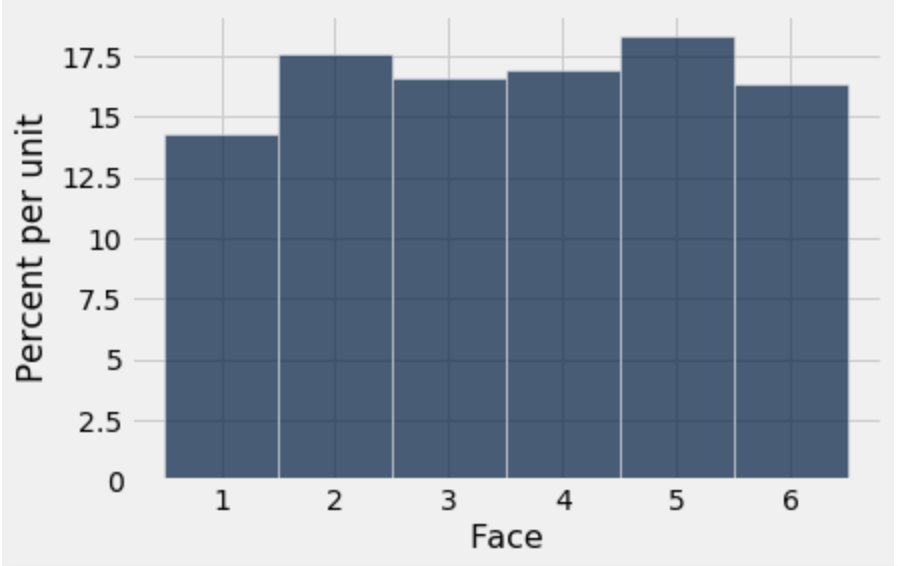
for i in \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_:

face = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

simulated\_rolls = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. We’ve generated histograms of dice roll results for samples of size 10, 100, 1000, and 10,000 below. Which histograms correspond to which sample sizes, and why?

A B

CD 

# 2 Hypothesis Testing

## Key Concepts

Suppose you flip a (presumably) fair coin 20 times, and see that the coin comes up heads 18 out of the 20 flips of the coin. This seems strange to you, as you previously believed that the coin is fair. A natural question to ask would be - was the 18 heads in 20 flips due to random chance? Or was it due to something other than random chance?

*Hypothesis testing* uses the power of computation to allow us to answer the question of “Was this strange event due to random chance?” in a scientific and consistent manner.

## Practice Problems

**2.1** Suppose you are flipping thumbtacks, and thumbtacks always either land pointing up or pointing down. You flip a thumbtack 60 times, and observe the thumbtack land pointing down 45 times. Your friend tells you that a thumbtack lands down with a 2/3 chance, and lands up with a 1/3 chance.

a. Does the thumbtack that you are flipping seem consistent with your friend’s model?

b. Complete the function flip\_thumbtack, which takes in no arguments and randomly flips a thumbtack 60 times. The thumbtack lands down with probability 2/3 and lands pointing up with probability 1/3. The function returns the number of pointing down results out of the 60 tosses.

def flip\_thumbtack():

probabilities = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

proportions = sample\_proportions(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

proportion\_down = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

return \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2.2** Suppose you want to leave your breakfast choices up to chance! You have a cabinet of 4 different cereal brands: Cheerios, Lucky Charms, Fruit Loops, and Cocoa Puffs. Suppose you randomly pick 4 cereal boxes *with replacement*.

a. What is the probability that you pick four unique brands of cereal?

b. What is the probability that you don’t pick Cheerios?

**2.3** In the Netherlands, all men take a military preinduction exam at age 18. The exam includes an intelligence test known as “Raven’s progressive matrices” and includes questions about demographic variables like family size. A study was done in 1968, relating the test scores of 18-year-old men to the number of their brothers and sisters. The records of all exams taken in 1968 were used.[[1]](#footnote-0)

1. What is the population of the study? What is the sample used in the study?
2. Is there a need to apply inference techniques to predict the mean score, max score, etc.? Why or why not?

1. Taken from Statistics Fourth Edition by Friedman, Pisani and Purves [↑](#footnote-ref-0)