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?

The probability that we roll a 5 is ⅙.   
This is a theoretical probability. We could get an empirical probability by simulating the experiment over and over again and estimating the theoretical probability based on the frequency of occurrence of the event we are interested in. We expect that the more times we simulate our experiment, the better we will be able to estimate the theoretical probability (hence our empirical probability will converge to the theoretical one). This is often referred to as the Law of Large Numbers.



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():

return np.random.choice(dice.column(‘Side’))

(Add second solution w/ .sample)

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 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

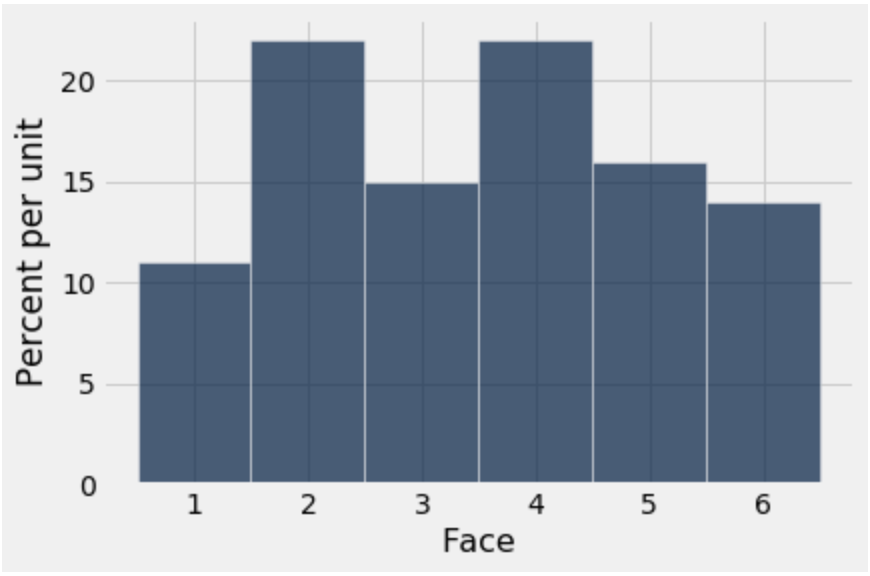
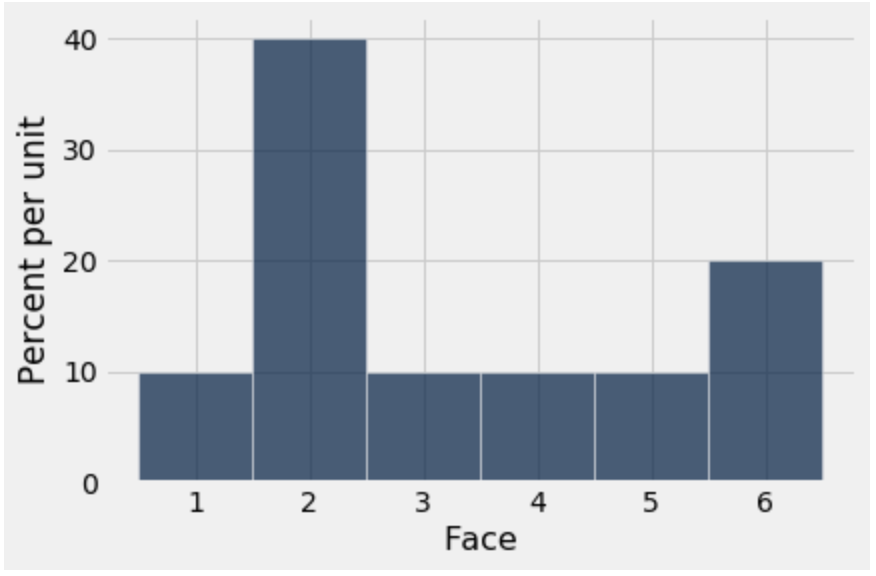
simulated\_rolls = make\_array()

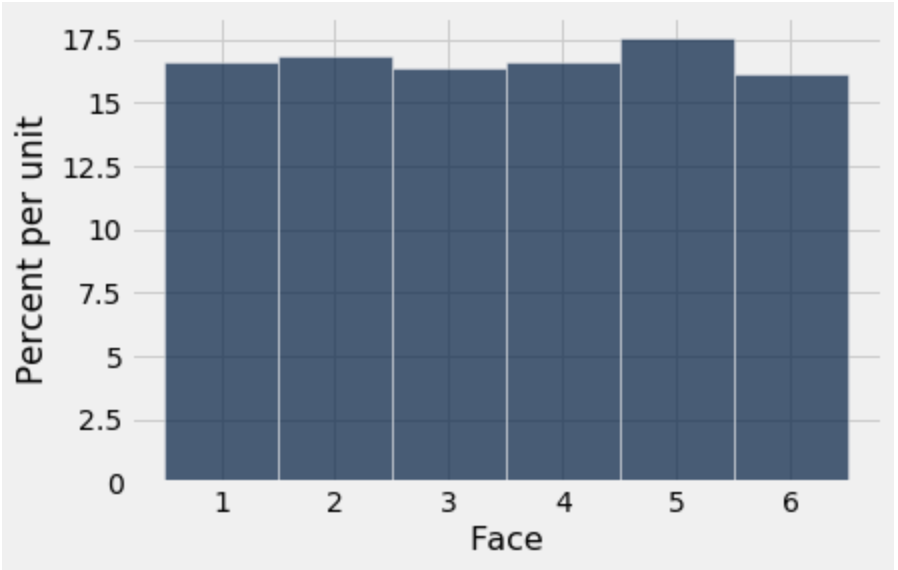
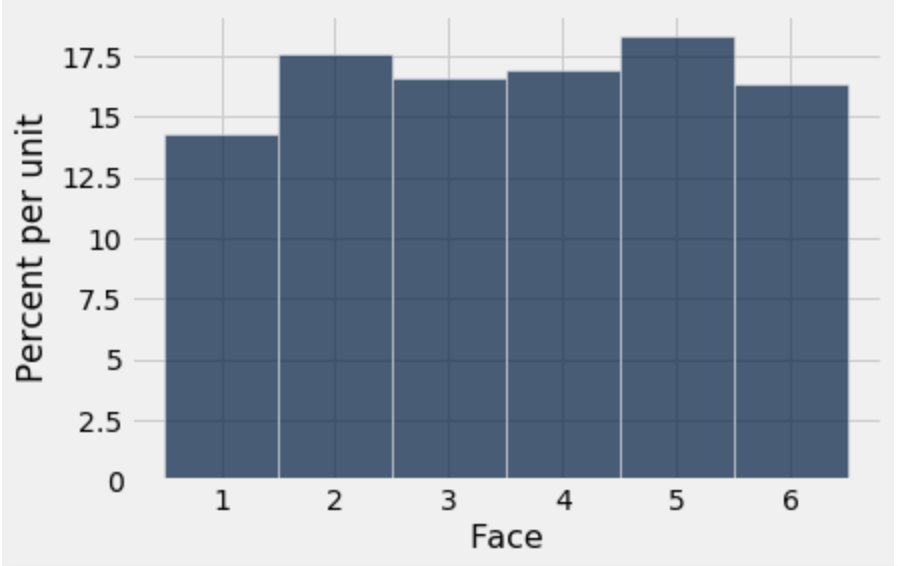
for i in np.arange(10):

face = roll\_die()

simulated\_rolls = np.append(simulated\_rolls, face)

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 

Answer: (A, 100), (B, 10), (C, 10000), (D, 1000)  
As the number of trials increases, the empirical distribution should converge to the probability distribution. Intuitively, you can think that in a small number of trials, there is a lot of randomness that might make the distribution diverge from what we would predict.

# 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 ⅔ chance, and lands up with a ⅓ chance.

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

Yes, the thumbtack you are flipping seems consistent with your friends model. The model stated a probability of ⅔ landing down, which would mean that we expect 40 thumbtacks to be landing down, but we observed 45, which is not too different from the expected amount.

b. Complete the function flip\_thumbtack, which takes in no arguments and randomly flips a thumbtack that lands down with ⅔ probability and lands pointing up with ⅓ probability 60 times, and then returns the number of pointing down results out of 60 tosses.

def flip\_thumbtack():

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

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

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

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

def flip\_thumbtack():

probabilities = make\_array(⅔, ⅓)

proportions = sample\_proportions(60, probabilities)

proportion\_down = proportions.item(0)

return proportion\_down \* 60

**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?

On the first draw, you can pick any of the four boxes, on the second pick you can only pick ¾ boxes, etc. Therefore,

P(Picking four unique brands of cereal) = (4/4)\*(3/4)\*(2/4)\*(1/4)

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

In order to not pick Cheerios, you have to pick Lucky Charms, Fruit Loops, and Cocoa puffs on all four draws. This probability is equivalent for all draws, since you are sampling with replacement.

Therefore, P(Not picking Cheerios) = (3/4)\*(3/4)\*(3/4)\*(3/4)

**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?

The population of interest to the researchers was all the 18-year-old men in the Netherlands in 1968. We know that all 18-year-old men take the test in the Netherlands and as a result, our sample is the population because all records were used.

1. Is there a need to apply inference techniques to predict the mean score, max score, etc? Why or why not?

There is no reason to apply any inference techniques. We have data on the whole population of interest (assuming no one skips the exam) and as a result we do not have to infer anything about the population.

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