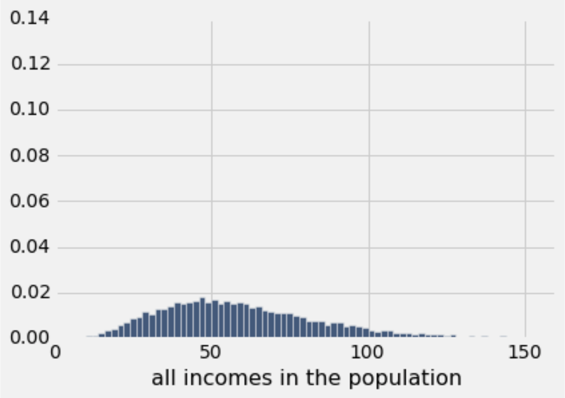
STAT/CS 94, FALL 2015 MIDTERM EXAM QUESTIONS A. Adhikari

1. The list **cards** contains credit card numbers. Each element is a 16-digit integer; for example, **1234567891011121** and not **"1234 5678 9101 1121"**.

a) Write a Python expression that evaluates to the total number of cards in the list.

b) Define a function **last_four** that takes a 16-digit integer card number as its input and returns the output **"Card ending in XYZW"** where XYZW are the last four digits of the card number. For example, **last_four(1234567891011121)** should return **"Card ending in 1121"**.

2. The histogram below represents data from a population of 10,000 incomes. The incomes are measured in thousands of dollars. Note that some bars on the left and right ends of the histogram are too small to be visible on the scale of the figure.



Page 2 of this exam contains six empirical histograms based on samples drawn uniformly at random with replacement from the population of incomes above.

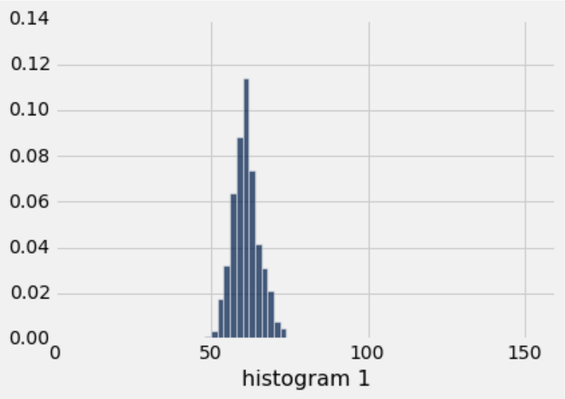
Match each description in the list below with the number of the appropriate histogram on Page 2. Explain your choices briefly.

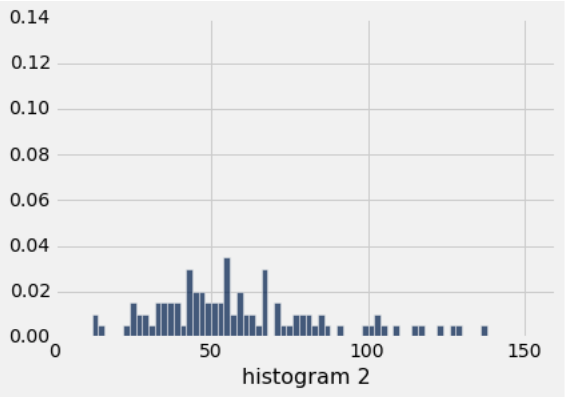
A: empirical distribution of the incomes in a sample of size 100

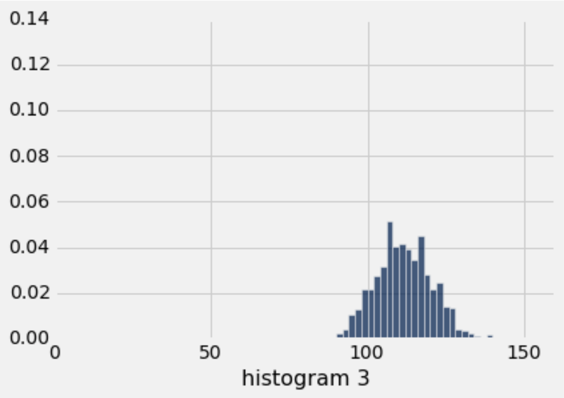
B: empirical distribution of the smallest income in a sample of size 100, based on 2000 replications of the sampling process

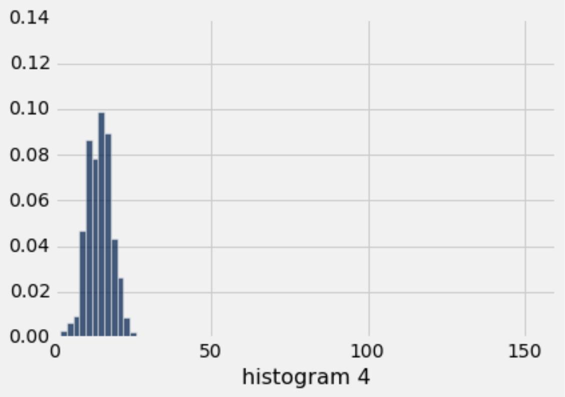
C: empirical distribution of the incomes in a sample of size 600

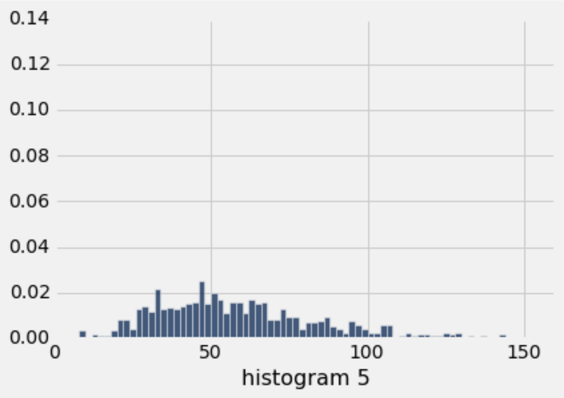
Empirical histograms for Problem 2

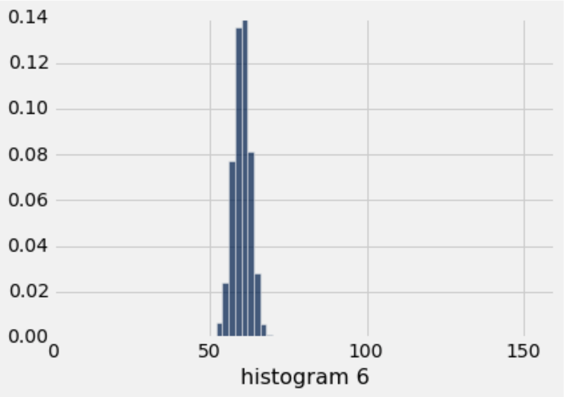




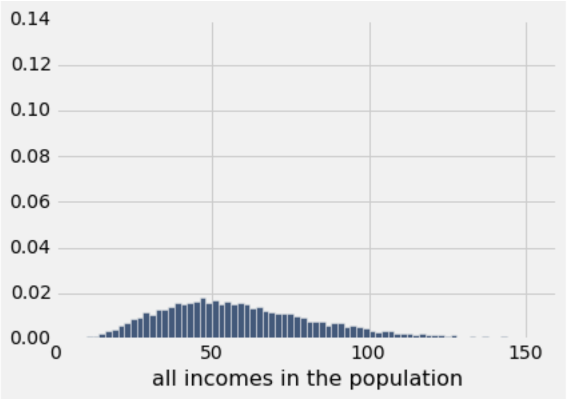








3. The histogram of the population of incomes in Problem 2 is displayed again here for ease of reference. It shows the distribution of 10,000 incomes measured in thousands of dollars.



a) The values of the incomes are contained in an array called **incomes**. The list of four numbers below consists of the results of the calls **np.median(incomes)**, **np.mean(incomes)**, and **np.std(incomes)**, as well as an irrelevant number, in scrambled order. Which is which, and why?

(i) 5.01 (ii) 26.71 (iii) 56.11 (iv) 60.07

b) The histogram of **incomes** is redrawn, still to the density scale but this time with different bins. Here are the bins and the heights of the bars. Show the calculation that will compute the missing height; leave all the arithmetic unsimplified.

bin (thousands of dollars)	0 – 20	20 – 40	40 – 70	70 – 100	100 – 200
height (proportion per thousand dollars)	0.0013	0.0108		0.0077	0.0015

4. The class Data Science 1A has several discussion sections. The instructor maintains her records in a table called **students**. The table has one row for each student in the class, and several columns. The column labeled **SID** contains the student ID number, and the column **year** contains the year in school (freshman, sophomore, junior, or senior). The GSIs also have a table with one row per student. This table is called **sections** and has many columns. The column **ID** contains the SID number, the column **section** contains the section number, and the column **Q4** contains the score on Quiz 4.

a) Write **one line** of code that constructs a table that has one row for every student, and columns **SID**, **year**, **section**, and **Q4** containing the corresponding data; your code should assign the table to the name **ds1a**. It is fine for the table to have more columns than the ones listed here.

b) Write **one line** of code that uses **pivot** to produce a table that has rows corresponding to the years and columns corresponding to the sections. For a given year and section, the body of the table should contain the highest Quiz 4 score of students in that year in that section. You are free to use any of the tables **students**, **sections**, and **ds1a** in your code.

5 . If possible, define a function **conclusion2** that takes one numerical input and returns the conclusion of any statistical test that has a P-value equal to the input and a cutoff of 2%. Specifically, for an input in the range $[0, 1]$, the function should treat the input as a P-value and return the appropriate one of the two strings **"the data support the null"** and **"the data don't support the null"** depending on the relation between the input and the cutoff. For an input outside the range $[0, 1]$, the function should return the string **"not a P-value"**.

If you feel that it is not possible to define such a function without knowing what kind of test is being performed, explain why not.

Questions 6-8 are based on the description below.

A cereal company places a small plastic toy inside each box of its cereal. The toy could be a bear, a ladybug, a car, or a dinosaur. Customers are encouraged to buy the cereal and collect these toys. If a customer has all four different kinds of toys, he/she gets a prize.

Assumptions of randomness. Each box is equally likely to contain any of the four kinds of toys, regardless of what is in the other boxes.

6. Suppose you buy four boxes of this cereal. Find the chance that you do not get the prize. Leave all the arithmetic unsimplified.

7. The calculation in Problem 6 is based on the assumptions about randomness that are stated above. To test these assumptions, suppose that you buy 100 boxes of the cereal, and get 27 bears, 20 ladybugs, 21 cars, and 32 dinosaurs. Use these data to set up a test of whether the assumptions are valid, as follows.

a) State the null hypothesis as a precise statement about chances.

b) Choose an appropriate test statistic (no explanations needed). Suppose you are going to perform a test of the null hypothesis by repeated sampling. Write **one line** of code that evaluates to **just one** value of your chosen test statistic generated by simulating the toys in 100 cereal boxes under appropriate assumptions.

8 (continuing Problem 7). Suppose you use the test statistic you chose in Problem 7b to perform a test of whether the data support the null hypothesis of Problem 7a.

Fill in the blanks with the appropriate choice from the list of terms below or with a numerical range. Examples of numerical ranges are "6.8 to 10.1," "10.1 or less," "more than 6.8," etc. No explanations are needed for your choices of terms, but if you use a numerical range, please show your calculations.

null alternative empirical distribution probability distribution

proportion probability test statistic exact approximate P-value

a) The P-value is the probability, assuming that the (i) hypothesis is true, of getting a (ii) in the range (iii) .

b) By running 3000 repetitions of simulating the toys in 100 cereal boxes, under the (iv) hypothesis, an (v) P-value can be obtained based on the (vi) of the test statistic.