Data 8 Spring 2016 hw06, Due: 5pm Tuesday, March 15

Please write your answers in the space provided. You can write on a printed copy or fill in the blanks with a PDF editor such as Acrobat Reader or Apple Preview. (Beware: some people have found that editing a PDF in a browser doesn't work.) When you're done, upload a scanned copy to Gradescope (`gradescope.com`). This assignment is due 5pm Tuesday, March 15. You will receive an early submission bonus point if you turn it in by 4pm Monday, March 14.

You are welcome to use `ds8.berkeley.edu` to try out Python expressions. Directly sharing answers is not okay, but discussing problems with course staff or students is encouraged.

Collaborators:

Problem 1 (Causality)

For each study below, describe the *treatment* variable, the *outcome* variable, and whether or not the study provides evidence of a *causal* relationship between the treatment and outcome.

(a) **(Study A: Soda)** A team analyzed data on daily soda intake and obesity from 622 participants. Each participant recorded how much soda they drank as well as their weight. The study found that people who drank soda regularly (i.e., they drank at least one can daily) were 33% more likely to be obese.

(b) **(Study B: Artifical Sweetener)** A study randomly assigned rats into two groups. They exposed each group of rats to the same amount of food, exercise, and living conditions, but fed one group artifically sweetened yogurt, and fed the other group unsweetened yogurt. They found that the first group had significantly more body fat on average after 12 weeks than the other group.

(c) **(Study C: Chocolate)** A study examined the eating habits of 8,000 men over 65 years, selected at random from the population of a particular state. They found that men who ate chocolate lived on average 1 year longer than those that did not eat any chocolate.

Study	Treatment	Outcome	Evidence of Causality? (yes/no)
Study A: Diet Soda			
Study B: Artifical Sweetener			
Study C: Chocolate			

Answer:

Study	Treatment	Outcome	Evidence of Causality? (yes/no)
Study A: Diet Soda	Drank soda regularly	Obesity rate	No
Study B: Artifical Sweetener	Artificial sweetener	Amount of body fat	Yes
Study C: Chocolate	Chocolate	Longevity	No

Problem 2 (Millenials)

A table called `people` contains two columns: a column of strings labeled `name` that contains names of people, and a column of integers labeled `age` that contains their current ages. Each name is unique (distinct from all other names).

(a) Write a Python expression whose value is the total number of people described by the table.

(b) Define a function called `millenial` that takes a person's `name`, and returns `True` if their age is between 16 and 36 (including either 16 or 36), and `False` otherwise.

(c) Add a new column called `is_millenial` to `people` where each row is True if the person is a millenial, and False otherwise. *Hint:* Use the `millenial` function you just wrote.

(d) Create a two-row table that contains in one of its columns the following two counts: the number of millenials and the number of non-millenials under age 50.

Answer:

(a)
```
people.num_rows
```

(b)
```
    def millenial (name):
        age = people.where(people.column("name")==name).column("age").item(0)
        return age >= 16 and age <= 36)
```

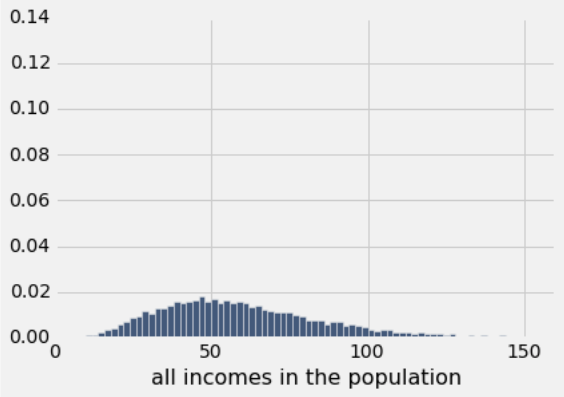
(c)
```
people.append_column("is_millenial", people.apply(millenial, "name"))
```

(d)
```
people.where(people.column("age") < 50).group("is_millenial")
```

Problem 3 (Histograms)

This question is taken from the Fall 2015 midterm. The vertical axis of each histogram is labeled in proportion per unit (so that total area sums to 1), rather than percent per unit as in class.

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



The next page contains six empirical histograms based on samples drawn uniformly at random with replacement from the population of incomes above.

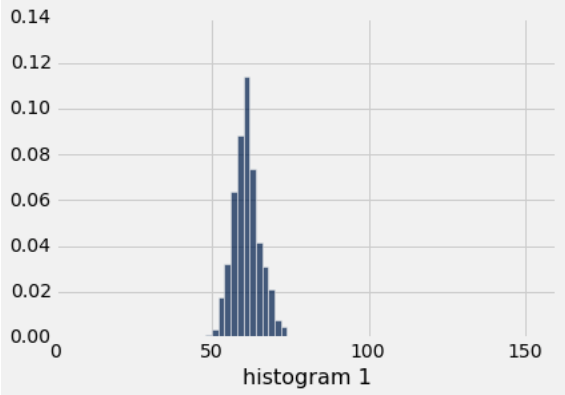
Match each description in the list below with the appropriate histogram on the next page. Explain your choices briefly.

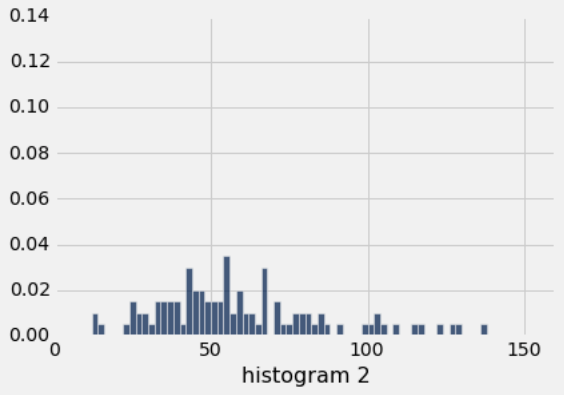
(a) empirical distribution of the incomes in a sample of size 100

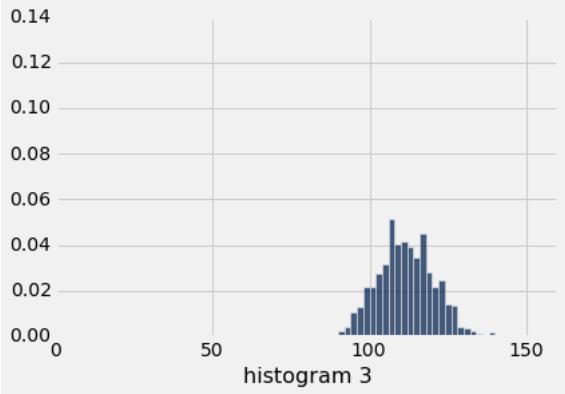
(b) empirical distribution of the smallest income in a sample of size 100, based on 2000 replications of the sampling process

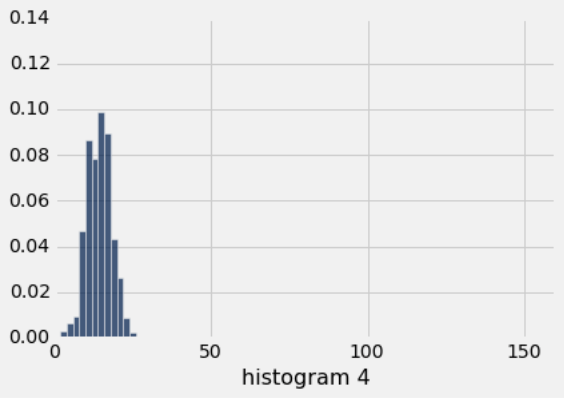
(c) empirical distribution of the incomes in a sample of size 600

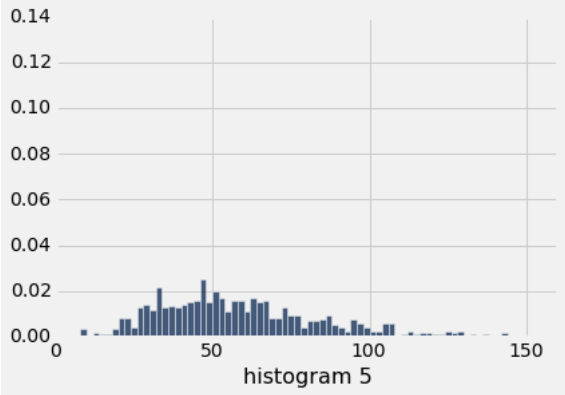
Empirical histograms for Problem 3

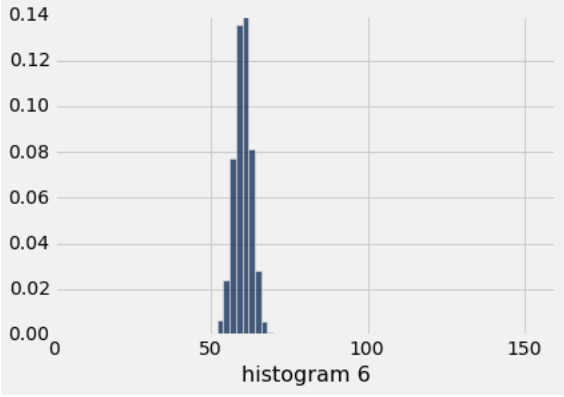












Answer:

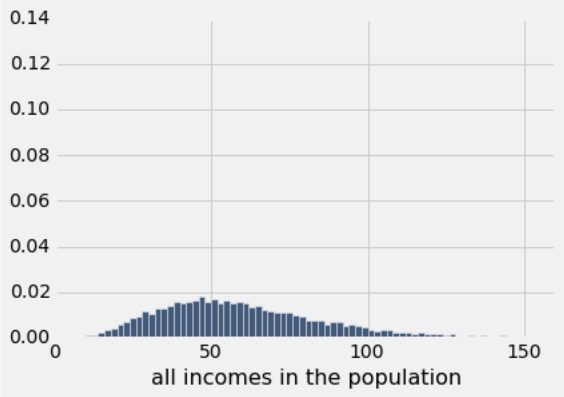
(a) histogram 2. similar shape to original histogram but not as close as histogram 5.

(b) histogram 4. the samples would likely have minimums that are small (e.g., less than 50), and histogram 4 is the only histogram that shows this.

(c) histogram 5. similar shape to original histogram and a closer match than histogram 2.

Problem 4 (Histograms continued)

The histogram of the population of incomes in Problem 3 is displayed again here for 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) Suppose that the histogram of `incomes` is redrawn, still to the density scale but this time with different bins. A table containing the new bins and heights of bars is given below. Show the calculation that will compute the missing height, leaving 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

Answer:

(a) (i) irrelevant
(ii) `np.std(incomes)` 5.01 is too small for the standard deviation (Chebyshev) and 56.11 is too large. the average distance from the mean could however be around 26.71.
(iii) `np.median(incomes)` 56.11 is around the middle but not pulled towards the skewed tail as greatly as the mean (60.07)
(iv) `np.mean(incomes)` 60.07 is around the middle and pulled a bit higher because of the right skew

(b)

$$20 * 0.0013 + 20 * 0.0108 + 30 * height + 30 * 0.0077 + 100 * 0.0015 = 1$$
$$\implies height = (1 - 20 * 0.0013 - 20 * 0.0108 - 30 * 0.0077 - 100 * 0.0015)/30$$

Problem 5 (Temperature)

In Homework 5, we worked with a dataset of temperature records for many places in the world. Now, let's look at the relationship between the temperature at a station on one day and the temperature at the same station a month later. Figure 1 is a scatter diagram displaying the temperature data for June 1 and July 1 in 2015. Here is how we made the graph:

- We put the June 1 and July 1 temperature data in standard units.
- We rounded the June 1 temperatures so that the data formed the vertical slices you see.
- Each red square is centered at the mean of its vertical slice. Its area is proportional to the number of data points in its slice.
- The blue line is the 45-degree line (where the June 1 temperature was equal to the July 1 temperature).

(a) Is the correlation between temperature on June 1 and July 1 positive or negative, or can you not tell from this graph?

(b) Would you guess that the magnitude of the correlation (that is, the absolute value of the correlation, $|r|$) is bigger than .95 or smaller than .95?

(c) Do you think the blue line is the least-squares regression line? Why or why not? If not, draw in your guess at the true line.

Answer:

(a) Positive. There's a clear upward trend.

(b) Smaller than .95. A correlation of .95 means the scatter plot looks much more like a line, as in Figure 2.

(c) No. For a graph of data in standard units, the slope of the best-fit line is the correlation. The 45-degree line has slope 1, but the correlation is clearly not 1. The true line has slope somewhat less than 1 and roughly passes through the means of the vertical strips. See Figure 3.

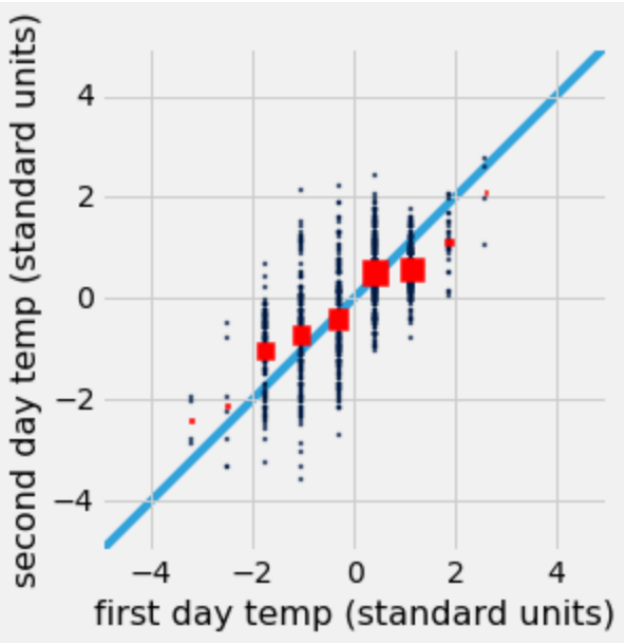


Figure 1: Temperatures on June 1 and July 1, 2015, at different weather stations.

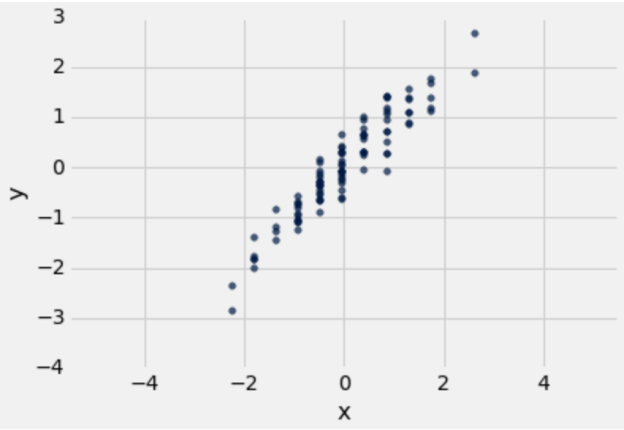


Figure 2: A scatter plot of a dataset with correlation .95.

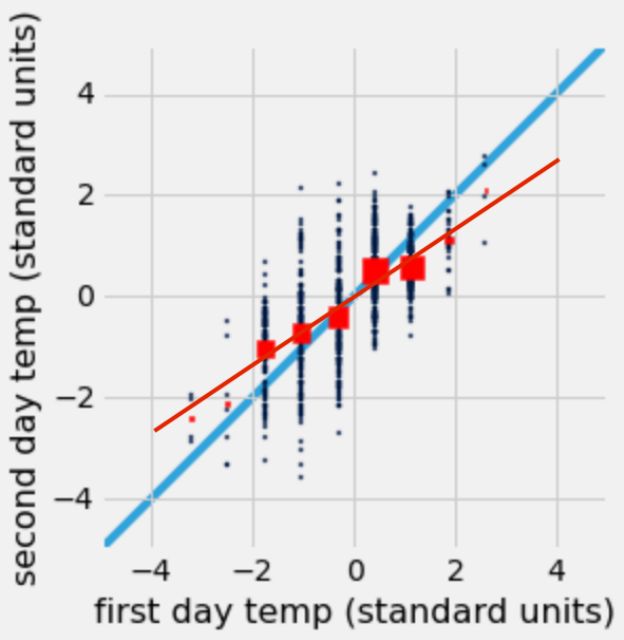


Figure 3: The true regression line.

Problem 6 (Analysis)

Each data analysis scenario below contains an error. Correct it by describing the change needed or rewriting the code correctly.

(a) We have a set of emails in a table called `email`, and we want to find the ones that are spam. There is one column called `text` that contains the text of each email. Suppose we want to examine all the emails that are in all-caps. One way to check whether a string is all-caps is to check whether it is equal to the upper-cased version of itself produced by the `upper` method. So we write this code to print out all the all-caps emails:

```
for email_text in email.column("text"):
    if email_text == email.upper():
        print("Potential spam email: " + email_text)
```

(b) We have a dataset about countries and their energy usage. It's a table named `energy` in which each country has one row, with one column named `pop` for the population of the country, and another column named `epp` for the amount of electricity used in the country **per person** in 2015. To figure out the amount of electricity used on Earth *per person* in 2015, we write:

```
np.average(energy.column("epp"))
```

(c) We are trying to figure out whether there is an association between the amount of foreign aid received by a developing country and its GDP growth rate. We have a table called `growth_rates` containing the name (in the column `country`) and growth rate in 2015 (in the column `growth rate`) of every country. We have another table called `aid` with one row for each time one country's government gave aid money to a developing country's government in 2015. (Some countries receive multiple aid packages each year, and each one is a separate row in this table.) `aid` has a column for the donor country's name (`donor`), a column for the recipient developing country's name (`recipient`), and a column for the amount of aid given (`amount`). We want to make a table called `growth_and_aid` containing both the growth rate and amount of money received by each country. So we join the two tables by calling:

```
growth_rates.join("country", aid, "recipient")
```

Answer:

(a) Instead of `email.upper()`, the second line should contain `email_text.upper()`

(b) This will compute the average per-capita electricity usage *across countries*. A country with a population of 1 will be treated in the same way as one with a population of 1 billion. To figure out the amount of energy per person, we have to somehow account for the population of each country. One way to do this is to compute the *total electricity usage* for each country, then the *total electricity usage* for the Earth, then divide by the total population of the Earth:

```
total_electricity_by_country = energy.column("epp") * energy.column("pop")
total_electricity = sum(total_electricity_by_country)
total_population = np.sum(energy.column("pop"))
total_electricity / total_population
```

(c) This isn't quite right. Each country's entry in `growth_rates` will be associated with *one* of the times it received aid, instead of its total aid. We need to first group the `aid` table by recipient country, summing up the aid received:

```
aid_totals = aid.group("recipient", sum)
growth_rates.join("country", aid_totals, "recipient")
```