STAT/CS 94 Fall 2015 Adhikari HW04, Due: 9/23/15

NAME: SID:

Please write your answers on the printed page, in the space provided. If you print this assignment yourself, please print it *double-sided* on a single sheet of paper. (There will be a small penalty for not following this instruction; it makes the grader's job more difficult.)

Problem 1 *Historic Histograms*

In a histogram of hours of employment of a group of students, the bins include the left endpoint but not the right. The table below shows two of the bins and the heights of the corresponding bars, which are in the *density* scale. (As a reminder, this is the kind of histogram a call to `.hist(..., normed=True)` produces.) If the two bins are combined into one, what would be the height of the bar over it?

hours	8 – 10	10 – 15
height	0.15 per hour	0.02 per hour

Answer: In a density histogram, the area of a bar equals the proportion of the graphed population in the bar's bin:

$$\text{Area} = (\text{Proportion in bin}).$$

The area of a bar is also, by definition, equal to its width times its height, or in other words the length of its bin's interval times its height:

$$\text{Area} = (\text{Bin length}) \times (\text{Bar height}).$$

Putting these together, we get:

$$(\text{Proportion in bin}) = (\text{Bin length}) \times (\text{Bar height}). \tag{1}$$

So the proportion in the first bin in the table is $(0.15/\text{hour}) \times ((10 - 8)\text{hour}) = .3$, and the proportion in the second bin is $(0.02/\text{hour}) \times ((15 - 10)\text{hour}) = .1$; the total proportion in both bins is .4. Now, the combined bars will cover the interval 8 to 15. Using equation 1 and solving for bar height, we have:

$$\begin{aligned}(\text{Bar height}) &= \frac{(\text{Proportion in bin})}{(\text{Bin length})}\\ &= \frac{.4}{15-8}\\ &= \frac{.4}{7}\\ &\approx .057\end{aligned}$$

Problem 2 *Zany Zips*

A table called `Contact_Info` consists of one row per person. The columns contain the person's name, email address, and other contact information. One of the columns is called Address and contains the person's street address written as a single string that ends with a zip code. For example, one of the entries is "1600 Pennsylvania Avenue NW, Washington, DC 20500" and another is "367 Evans Hall #3860, University of California, Berkeley, CA 94720-3860".

(a) Write a function called `zip_code` that takes as its argument an address string such as the examples above, and returns the zip code as written in the address string. Thus `zip_code(‘1600 Pennsylvania Avenue NW, Washington, DC 20500’)` should return `‘20500’` and `zip_code(‘367 Evans Hall #3860, University of California, Berkeley, CA 94720-3860’)` should return `‘94720-3860’`.

(b) Write code that augments the table `Contact_Info` with a column called ‘Zip Code’ that contains the zip codes extracted from the Address column by the function `zip_code` that you created in part (a). Again, your code should cause the ‘Zip Code’ column to be *added* to the table named `Contact_Info`.

Hint: Since you probably didn't write your function to also work on NumPy arrays, you may want to use the `Table` method `.apply()`. Look up the syntax at data8.org/datascience/tables.html if you need to.

Answer:

(a) There are many ways to write this function. If we were writing it ourselves we would use the String method `.split(’ ’)`, which makes it easy. Let's use conditionals, as we had just learned in the class.

It's actually not too complicated. We have been told that the address always ends with a zip code. Zip codes are at least 5 characters long, so it's always okay to check the 5th-from-the-last character. (If `address` might not contain a zip code, or might not be 5 characters long, then we'd have to think about what to do in those cases. Fortunately, we don't. Note that real data rarely provides ironclad guarantees, so in practice it would be a good idea to make your code more robust than ours.) There are actually just two cases: the zip code is of the form “XXXXX-XXXX”, or it is of the form “XXXXX”. We can distinguish between the two just by checking for a dash in the 5th-from-the-last character.

```
def zip_code(address):
  if address[-5] == ’-’:
    return address[-10:] # Shorthand for address[(len(address)-10):len(address)]
  else:
    return address[-5:] # Shorthand for address[(len(address)-5):len(address)]
```

Though it may be possible to reason perfectly about short pieces of code, most of us aren't very good at it, so it's always a good idea to test your code. In our case, we tested `zip_code("foo 94704")`, `zip_code("94704")`, `zip_code("94704-1234")`, and `zip_code("foo 94704-1234")`.

Note also that zip codes are not numbers, since sometimes they contain dashes. So `zip_code` should leave its return value as a string, rather than try to convert it to a number.

(b) `Contact_Info["Zip Code"] = Contact_Info.apply(zip_code, "Address")`

Problem 3 *School Shuffle*

The Registrar's Office at a school has the school's complete enrollment information. For every student, it has the complete list of classes in which the student is enrolled, and for every class, it has the complete list of enrolled students. There are 300 classes being offered at the school. The Registrar selects a probability sample of students as follows: she selects a sample of 10 of the 300 classes at random without replacement, and then for each selected class, she adds all the students in that class to the probability sample. (So if a student is in more than one of the selected classes, that student appears in the probability sample more than one time.) This method is called “cluster sampling” because each selected class is a cluster of students.

(a) Do all students have the same chance of entering the sample? Explain your answer.

(b) Describe three significant differences between the properties of the Registrar's sample and properties of a sample drawn at random without replacement from among all the students.

Answer:

(a) No. In this sampling scheme, a student has one chance to enter the sample for each class she is enrolled in, so students enrolled in more classes have a higher chance of being in the sample. (Note that students can appear more than once in the sample, and students enrolled in more classes also *appear more times on average* in the sample.)

You didn't need to do this, but we can compute the probability that two example students enter the sample. The computation is similar to the computation we did for the birthday problem. Say that Alejandro is in 3 classes and Barbara is in 6 classes. The event that Alejandro doesn't enter the sample is the event that he isn't in the first sampled class, and isn't in the second sampled class, $\cdots$, and isn't in the tenth sampled class. Imagine choosing the sampled classes one at a time. Since the *classes* are selected without replacement, each time one of Alejandro's classes isn't selected, the total number of remaining classes goes down, and so does the number of classes not containing Alejandro. So the probability Alejandro doesn't enter the sample is:

$$\frac{(300-3)}{300} \times \frac{(299-3)}{299} \times \frac{(298-3)}{298} \cdots \times \frac{291-3}{291} \approx .903.$$

Similarly, the chance that Barbara doesn't enter the sample is

$$\frac{(300-6)}{300} \times \frac{(299-6)}{299} \times \frac{(298-6)}{298} \cdots \times \frac{(291-6)}{291} \approx .815.$$

.

(b)

- A sample drawn at random without replacement (which is shorthand for a sample drawn *uniformly* at random without replacement) will contain each student with equal probability. Above we saw that the cluster sampling scheme gives a higher probability to students who take more classes.
- Students can appear in the cluster sample more than once – if two of a student's classes are drawn, she will be listed in the sample twice. A sample drawn without replacement, of course, contains only one or zero copies of each student.
- The size of the sample drawn at random without replacement is fixed (we have to decide it ahead of time). The size of the cluster sample is random; if we happen to draw several large classes, then our sample will be big.
- The probability that a student appears in the sample depends differently on whether another student is in the sample. (More precisely we might say that the *joint distribution* of two or more students is different in the cluster sample.)

  Say that we draw 50 students in the simple random sample without replacement (abbreviated SRSWOR from now on). In the SRSWOR, if we observe that the first student in the sample is Casey, then the chance that student Delilah is also in the sample always goes down from 50/300 to 49/299, since Casey has taken up one spot. In the cluster sample, this chance (called a *conditional probability*) depends on the classes Casey and Delilah are taking.

  Say that Casey only takes classes Delilah is in. Then, in the cluster sample, if we see that Casey is the first student in the sample, then we know Delilah also appears in the sample with probability 1, since one of Casey's (and hence Delilah's) classes must have been selected! On the other hand, if Casey wasn't in any of Delilah's classes, then Delilah's chance of appearing in the cluster sample goes down (to some number that is a bit harder to calculate) when we observe that Casey is in it.

Problem 4 *Casino Conundrum*

There are 38 pockets in a Nevada roulette wheel: 18 red numbers, 18 black numbers, and 2 green numbers. A bet on a "split" (two pockets–you win if either one is picked) pays 17 to 1.

(a) If you bet repeatedly on a split, then in the long run roughly what proportion of bets do you expect to win? What proportion do you expect to lose?

(b) If you repeatedly bet \$1 on a split, then every bet you win gives you a net gain of \$17, and every bet you lose gives you a net gain of −\$1. In the long run, what do you expect as your average net gain per bet? [If you have trouble figuring this out, first work with 5 bets and pretend that the sequence of results was Lose, Lose, Lose, Win, Lose. Figure out your average net gain per bet, in terms of the proportions of wins and losses.]

(c) Repeat part (b) if instead you are betting \$1 each time on "red"; the bet pays 1 to 1.

(d) Repeat part (b) if instead you are betting \$1 each time on a "single number" (one pocket); the bet pays 35 to 1.

Answer:

1. Betting on a split wins 2 out of 38 times, so the probability of winning is $\frac{2}{38}$. In the long run, the "law of averages" says that the proportion of bets we expect to win equals the probability of winning, so we expect to win $\frac{2}{38}$ of our bets in the long run. Similarly, the probability of losing is $\frac{36}{38}$, and we expect to lose $\frac{36}{38}$ of our bets in the long run.

2. Say that we bet 1,000,000 times on a split. Let X be the number of times we win and Y be the number of times we lose. Note that X and Y are *random.* Then our total winnings will be $17 \times X + (-1) \times Y$; let's denote that random quantity (our total winnings after 1,000,000 bets) by Z. We don't have a law of averages for X or Y, meaning that we don't think X will be really close, in absolute terms, to its average value. (It's not so improbable that 50 bigger than its average value, for example.) But the law of averages does say that $\frac{X}{1000000}$, the *proportion* of wins, will be very close to its average value, which is $\frac{2}{38}$. So, we have:

$$\begin{aligned} \frac{Z}{1000000} &= 17 \times \frac{X}{1000000} + (-1) \times \frac{Y}{1000000} && (2) \\ &\approx 17 \times \frac{2}{38} + (-1) \times \frac{36}{38} && (3) \\ &= \frac{-2}{38}, && (4) \end{aligned}$$

   where the ≈ comes from the law of averages.

   Once the concept is clear, we can eliminate the thought experiment with the 1000000 bets and just say directly that the long-run average winnings are:

$$17 \times \text{Prob(win)} + (-1) \times \text{Prob(lose)} = \frac{-2}{38}$$

   .

   Note that you don't need to answer questions in such detail, but it is generally a good idea to show your work.

3. Now the chance of winning is $\frac{18}{38}$. So, by the same reasoning as above, our long-run average winnings are:

$$1 \times \text{Prob(win)} + (-1) \times \text{Prob(lose)} = \frac{18}{38} - \frac{20}{38} = \frac{-2}{38}$$

4. Now the chance of winning is $\frac{1}{38}$, and our long-run average winnings are:

$$35 \times \frac{1}{38} + (-1) \times \frac{37}{38} = \frac{-2}{38}$$