ORCA 4500 HW5
Yihang Ding, yd2459
yd2459@columbia.edu

1. 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.

No, because the probability for a student to not enter the sample is the same as the probability to select 10 classes from which this student do not attend, as the number of classes a student attends varies, the probability changes, so not every student have the same chance.

(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.

1. The Registrar's sample depends on the distribution and the number of classes every student selects. But sampling from randomly selecting student each time is a Bernoulli distribution.

2. There might be duplicate students in the Registrar's sample, as students choose multiple classes, but as for the second sample, the students are drawn without replacement, there are no duplicate students in the sample.

3. There might be duplicate classes in the second sample, but no duplicates in the registrar's sample, the reason is the same as above.

2. In a computer dating application, each person answers 1000 questions, and on each question, the answer is an integer between 0 and 10. We say that two people match on an answer if their scores are within two of each other. We say two people are compatible if their answers match on at least k of the questions. Suppose that you have 100 people using your application and you want to be confident that with probability .99 at least one match exists. How should you set k?

Computation functions are in the Jupyter notebook.

Event A: two people match on a question

Event B: two people are compatible

Event C: at least 1 match exists

$P(A) = \frac{3\times2+4*2+5*7}{11\times11} = \frac{49}{121}$

The number of matches for two persons has a Bernoulli Distribution, $B\sim(1000, P(A))$

$P(B) = P(\text{num of matches} > k)$

$P(\text{not } B) = 1 - P(B)$

$P(C) = 1 - P(\text{not } C) = 1 - P(\text{not } B)^{\left|\frac{100}{2}\right|} > 0.99$

From P(C) we get P(not B)=0.9990

As $P(\text{not } B) = \sum_0^k \left|\frac{1000}{k}\right| P(A)^k P(not\ A)^{1000-k} < 0.999$, we get k=454

So K should be set as 454.

3. When you flip coins, a streak of length k is a series of K consecutive flips that all come up with the same value. For example if your flips are HHTTHTTTHTHT, then there is one streak of length 3. The longest streak in a series of flips is the streak of greatest length, so in the sequence above, the longest streak is of length 3 and in the seqence THHHHTHHHTHHT the longest streak is of length 4. I want to understand what the expected length of the longest streak is, when I flip n coins. Perform some simulations to answer this question as best you can. Concretely, answer the question for n = 10, 100, 1000, 10000, 100000. If you can, try to come up with a general formula.

The experiment function is in the Jupyter notebook.

For each n, by doing experiments for 10000 times, get the following answers.

```
streak=test(10)
print('n=10, expected max streak length is:',streak)
```

```
n=10, expected max streak length is: 2.6699
```

```
streak=test(100)
print('n=100, expected max streak length is:',streak)
```

```
n=100, expected max streak length is: 5.9667
```

```
streak=test(1000)
print('n=1000, expected max streak length is:',streak)
```

```
n=1000, expected max streak length is: 9.2746
```

```
streak=test(10000)
print('n=10000, expected max streak length is:',streak)
```

```
n=10000, expected max streak length is: 12.6158
```

```
streak=test(100000)
print('n=100000, expected max streak length is:',streak)
```

```
n=100000, expected max streak length is: 15.9214
```

4. Given the following statistics, what is the probability that a woman over 50 has cancer if she has a positive mammogram result?

(a) One percent of women over 50 have breast cancer.

(b) Ninety percent of women who have breast cancer test positive on mammograms.
(c) Eight percent of women will have false positives.

P(A) = P(woman over 50 has cancer)
P(B) = P(woman has positive result)

With (a), P(A) = 1/100=0.01, P(not A)=0.99
With (b), P(B|A) = 0.9
With (c), P(B| not A) = 0.08

P=P(A|B)=P(AB)/P(B)=P(A)P(B|A)/P(B)

$$= \frac{P(A)P(B|A)}{P(AB)+P(not\ AB)} + \frac{P(A)P(B|A)}{P(B|A)P(A)+P(B|not\ A)P(not\ \mathrm{A})} = 0.1$$