DCCS 221 – Probability and Statistics

Lab 1

Data Preparation

1. Go to blackboard and download data file(SFBAdrivers.csv) to your DCCS221/Lab1 folder.

2. Open Python Editor and set your 'Work directory' to your Lab1 folder.

3. Import pandas library:

```
import pandas as pd
```

   Your data contains 4 columns as shown below. First column is the number of vehicles of the household (HHVEHCNT), and second is the number of drivers (DRVRCNT) and the column 4 is the household size (HHSIZE).

```
df.head
✓ 0.0s
<bound method NDFrame.head of      HHVEHCNT  DRVRCNT  HHSIZE
0           2        2       6
1           1        1       1
2           3        4       5
3           2        2       2
4           3        3       3
..        ...      ...     ...
582         2        2       2
583         6        2       2
584         1        2       2
585         3        2       2
586         1        1       1

[587 rows x 3 columns]>
```

4. Now let's deal with the dataframe (df) which created in your working environment (Check Variables).

Problems

1. Sample Space

   As mentioned in lecture, each variable has a sample space. In a nutshell, the empirical sample space is what actually appears in the data, and the theoretical sample space is what the values could conceivably be not limited to observations in the data.) We assume that the empirical sample space is defined as all integer values between the minimum value and the maximum value inclusive.

   a. What is the sample space of household vehicles?

      i. Empirical sample space: Python code: **min(df.HHVEHCNT), max(df.HHVEHCNT)**

      ii. Theoretical sample space: **( )**

   b. What is the sample space of household drivers?

      i. Empirical sample space: **min(df.DRVRCNT), max(df.DRVRCNT)**

      iii. Theoretical sample space: **( )**

   c. What is the sample space of number of persons per household?

      i. Empirical sample space: **min(df.HHSIZE), max(df.HHSIZE)**

      iv. Theoretical sample space: **( )**

   d. What is the size of the total empirical sample space, considering all three variables? Note that certain combinations of values are logically impossible. **Big Sample space – can't logically have more drivers than household members, but other than that, nothing else is logically impossible.**

2. Events

   Write out the following events in set notation (unions, intersections) and calculate the number of observations using Python. (*NOTE that this code has two equal signs. There should be NO SPACE between the equals sign. The space appears here so you can visually see that there are two.)*

   a. Household size is 2.

      i. Event in words: ***Event that HHVEHCNT== 2***

      ii. Python: ***len(df[df.HHVEHCNT == 2])***

   b. Number of drivers in the household is 2 and household size is 2.

      iii. Event in words: **( )**

      iv. Python: ***len(df[(df.DRVRCNT == 2) & (df.HHSIZE == 2)])***

   c. Number of vehicles in the household is 2 and number of household drivers is 2.

      i. Event in words: **( )**

      ii. Try the code here for yourself. **( )**

   d. Household size is larger than 5.

      i. Event in words: **( )**

      ii. Python: **( )**

   e. Household size is less than 3 and number of household vehicles is 3 or more.

      i. Event in words: **( )**

      ii. Python: **( )**

   f. Number of household drivers is 0 OR household vehicles is 0.

i. Event in words: ( )

ii. Python: ( )

3. Set Operations

Beyond finding an observation count for each event, we can create a vector of the actual observation number which meets the event criteria. For example, to see the observation number of the event the number of persons per household is 2, the event can be found with conditional expression (**HHSIZE ==2**). The following events are defined (if the definition is missing, fill it in yourself):

a. Event A: **number of persons per household is 2 or fewer**

b. Event A^c: ( )

c. Event B: **household vehicles is 0**

d. Event B^c: ( )

e. Find $A^c \cap B^c$ numerically using Python.

i. Create the vector of observation numbers which correspond to $A^c \cap B^c$. Also, name this vector p. *Hint:* ***p = df.query(event_Ac + ' & ' + event_Bc).index***

f. Find A ∪ B numerically using Python.

i. Create the vector of observation numbers which corresponds to (A ∪ B), and call it r. Hint: ***df.query(event_A + ' | ' + event_B)***

ii. Visually inspect the vector **p**, and compare it with the vector r. Pick an observation number in r and see if you can find it in vector **p**. Can you? Why or why not?

iii. Write Python code to perform the previous step. The code should take two lists of observation numbers and identify the numbers they have in common.

g. Find $(A \cup B)^c$ numerically, by subtracting the value of the number of observations found in (f) from the total number of observations, 587.

h. What do you notice about (g), and (e) & (f)?

i. Write out, using set notation and the symbols for events shown in (a-d), deMorgan's law as it applies in this case.

4. Probability – **Challenge Question (Extra pts.)**

Define each of the events in (a) through (f) in Problem 3 as its corresponding letter. For example, event a is number of persons in household is 2. Let $P(E)$ be the fraction of observations associated with event E.

a. Write Matlab code to evaluate $P(a)$, $P(b)$, $P(c)$, $P(d)$.

b. For each of the pairs of events drawn from a,b,c, and d—a&b, a&c, b&c —use Python code to demonstrate that:

i. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

ii. $P(E_1 \cap E_2) \geq P(E_1) + P(E_2) - 1$