Homework #3

Le Wang

Instruction: Do all the following empirical exercises using R. Turn in your R markdown file with answers and supporting tables and graphs, if any. Refer to the R output whenever appropriate when discussing your results.

Question 1 (Practice Question): [Sample Space, Events, and Naive Probability]

This question is similar to the examples in class, this type of question is probably redundant, as suggested by Carson! So, I leave it as a practice question only.

We will roll six dices. Answer the following questions for this experiment.

1. Using `R` to generate the sample space for rolling six dices. Show 10 outcomes from the sample space.
2. Using `R` to generate the event when the sum of six dices is greater than 30. Show the outcomes in this event.
3. Using `R` and the naive definition of probability to calculate the probability of the event in 2.

Question 2: [Counting and Sample Space]

There are 1000 fans on the wait list for the OU-Texas game tickets, but there are only 100 tickets left. The OU ticket office determines who will get a ticket through lottery. What is the likelidhood that both my wife and I would win the lottery? Note that this is slightly different from the simple example that I talked about in class, but don't worry, the logic is the same. Use `R` to answer each of the following question.

1. How many possible combinations of the people who would win the lottery?
2. How many possible combinations that my wife and I would win the lottery?
3. What is the probability that *both* my wife *and* I would win the lottery?

Question 3: [Counting, Sample Space and Naive Probability: The Birthday Problem Redux]

Many of you probably have seen this problem in your undergrad statistics course. And now lets see if we could solve it in `R`.

Suppose that there are $N = 30$ people together at a party. What is the probability of at least one pair of attendants with the same birthday? This problem tests your understanding of naive probability, some properties of probability (derived from the three axioms discussed in class), and counting techniques. Let me walk you through this.

We will ignore leap years and assume that there are only 365 days in a year. We will also assume that births are equally distributed over the course of a year. Let's first define the event A

$$A = \{\text{there is at least one match}\}$$

We are interested in $\mathbb{P}[A]$

1. What is the sample space for this experiment? **Hint:** Note that you have $N = 30$ people, and every one of them have 365 possibilities for his/her birthday.
2. What are the possible combinations for no match at all? **Hint:** You can think of this as a sequential experiment (someone is waiting at the party door to make sure that nobody with the same birthday as the previous person would come in). In other words, there are 365 possibilities for the first person, 364 possibilities for the second and so on.

3. What is the probability of the occurrence of a no-match party? **Hint:** Use the naive definition probability
4. What is the probability of the occurrence of event A? **Hint:** Use the property for probability for a complement of a set.

Question 4. [Further exploration of Law of Total Probability]

The law of total probability is a very important tool that links the probability for the entire population to the probabilities for sub-populations. Let's practice it again with a dataset called `mpg` that comes with the package `ggplot2`. We need to load the package `ggplot2` first.

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

Now let me tell you the probabilities with `year` (**year of manufacturing**) and `model` (**model name**).

```
mytable<-table(model=mpg$model,
               year=mpg$year)
dist <- as.matrix(prop.table(mytable))
dist
```

```
##                         year
## model                          1999        2008
##   4runner 4wd           0.017094017 0.008547009
##   a4                    0.017094017 0.012820513
##   a4 quattro            0.017094017 0.017094017
##   a6 quattro            0.004273504 0.008547009
##   altima                0.008547009 0.017094017
##   c1500 suburban 2wd    0.004273504 0.017094017
##   camry                 0.017094017 0.012820513
##   camry solara          0.017094017 0.012820513
##   caravan 2wd           0.025641026 0.021367521
##   civic                 0.021367521 0.017094017
##   corolla               0.012820513 0.008547009
##   corvette              0.008547009 0.012820513
##   dakota pickup 4wd     0.017094017 0.021367521
```

```
##   durango 4wd            0.012820513 0.017094017
##   expedition 2wd         0.008547009 0.004273504
##   explorer 4wd           0.017094017 0.008547009
##   f150 pickup 4wd        0.021367521 0.008547009
##   forester awd           0.008547009 0.017094017
##   grand cherokee 4wd     0.008547009 0.025641026
##   grand prix             0.012820513 0.008547009
##   gti                    0.012820513 0.008547009
##   impreza awd            0.017094017 0.017094017
##   jetta                  0.021367521 0.017094017
##   k1500 tahoe 4wd        0.008547009 0.008547009
##   land cruiser wagon 4wd 0.004273504 0.004273504
##   malibu                 0.008547009 0.012820513
##   maxima                 0.008547009 0.004273504
##   mountaineer 4wd        0.008547009 0.008547009
##   mustang                0.017094017 0.021367521
##   navigator 2wd          0.008547009 0.004273504
##   new beetle             0.017094017 0.008547009
##   passat                 0.017094017 0.012820513
##   pathfinder 4wd         0.008547009 0.008547009
##   ram 1500 pickup 4wd    0.012820513 0.029914530
##   range rover            0.008547009 0.008547009
##   sonata                 0.017094017 0.012820513
##   tiburon                0.008547009 0.021367521
##   toyota tacoma 4wd      0.017094017 0.012820513
```

Now, using this piece of information, calculate the percentage of cars made in year 2008 and 1999, respectively.

Question 5. [Probability Mass Function and CDF]

Again, use the `mpg` data to answer to the following questions.

1. Using the variable `class` (**vehicle class**), find out which type of vehicle is the most popular with two different approaches
    1. Count the actual number of car for each type.
    2. Plot the histogram of the variable with **probability**, as opposed to **counts**
2. Using the variable `cyl` (**number of cylinders**), find out the percentage of cars with an engine smaller than a 6-cylinder one. Note that this is nothing but your CDF for this variable.