Basava R Kanaparthi (110479710)
name

Probability and Statistics for Data Science
Stony Brook University
CSE594 - Spring 2016

Assignment 1: Probability Theory
Assigned: 2/3/2016; Due: 2/16/2016, 3:45pm EST

Part I. Problem Solving

1. **Combinatorics Touchup:** Consider the rolling of 10 (6-sided, fair) dice.
   a. **describe the sample space in words:**

      Ω = { $<D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 D_{10}>$, $D_i \in \{1,2,3,4,5,6\}$}
      Sequences of length 10, where each element can take values between 1 and 6 (inclusive).

   b. **How many possible outcomes are in the sample space? Why?**

      Each dice can take 6 possible values and each is independent with the others. Therefore, total number of possibilities is 6*6*6*6*6*6*6*6*6*6 = 6^{10}

   c. **What is the probability of rolling 7 sixes? (show your work)**

      Assuming all the 6 outcomes are equally probable, we define:

      $$p(6) = \frac{1}{6}$$

      $$p(!6) = \frac{5}{6}$$

      Using the binomial distribution:

      $$P(X = 7) = \binom{10}{7}\left(\frac{1}{6}\right)^7\left(\frac{5}{6}\right)^3$$

      Therefore, probability of rolling six dices = 0.000248

   d. **Suppose (given) we ended up with exactly 7 sixes after rolling all 10 dice, but some of the dice fell onto a chair and others onto the floor. What is the probability that all 7 sixes remained on the Table? (show your work, list assumptions)**

      Assuming all three locations, chair, table and floor are equiprobable (1/3).

A = All 7 sixes are on table. And since, the remaining 3 dices can be anywhere on the floor, table or chair, we are exhausting their sample space, i.e.,
P(A|B) over all (B = b) = P(A)
Therefore,

P(all seven 6's on table) = $\binom{7}{7}\left(\frac{1}{3}\right)^7$

$$= \left(\frac{1}{3}\right)^7$$

2. **Independent Offspring?** Suppose a family has three children.
   a. **What is the probability that all three are girls?**
   Each child's gender is independent of the other two.

$$P(GGG) = \left(\frac{1}{2}\right)^3$$

   b. **Suppose you learn that at least one child is a girl. Given this information, what is the probability that all 3 are girls? (show your work)**

$$A: At\ least\ one\ child\ is\ a\ girl\ \ !A: None\ is\ a\ girl\ (all\ boys)$$
$$B: All\ 3\ are\ girls$$
$$P(A) = 1 - P(!A) = 1 - (1/2)^3 = 7/8$$
$$P(B) = (1/2)^3 = 1/8$$
$$P(AnB) = 1/8$$

   Using Bayes theorem:

$$P(B|A) = \frac{P(AnB)}{P(A)}$$
$$= 1 / 7$$

   c. **Suppose you learn that the oldest child is a girl. Given this information, what is the probability that all 3 are girls? (show your work)**

$$A: Oldest\ child\ is\ a\ girl$$
$$B: All\ 3\ are\ girls$$
$$P(A) = ½$$
$$P(B) = 1/8$$
$$P(AnB) = 1/8$$
$$P(B|A) = \frac{P(AnB)}{P(A)}$$
$$= 1 / 4$$

   d. **Another family has 3 children (ignore all information above). You learn that only one of these children is a boy. Your friend knows which child it is and asks you to guess. You guess the oldest child. What is the probability that you are right?**

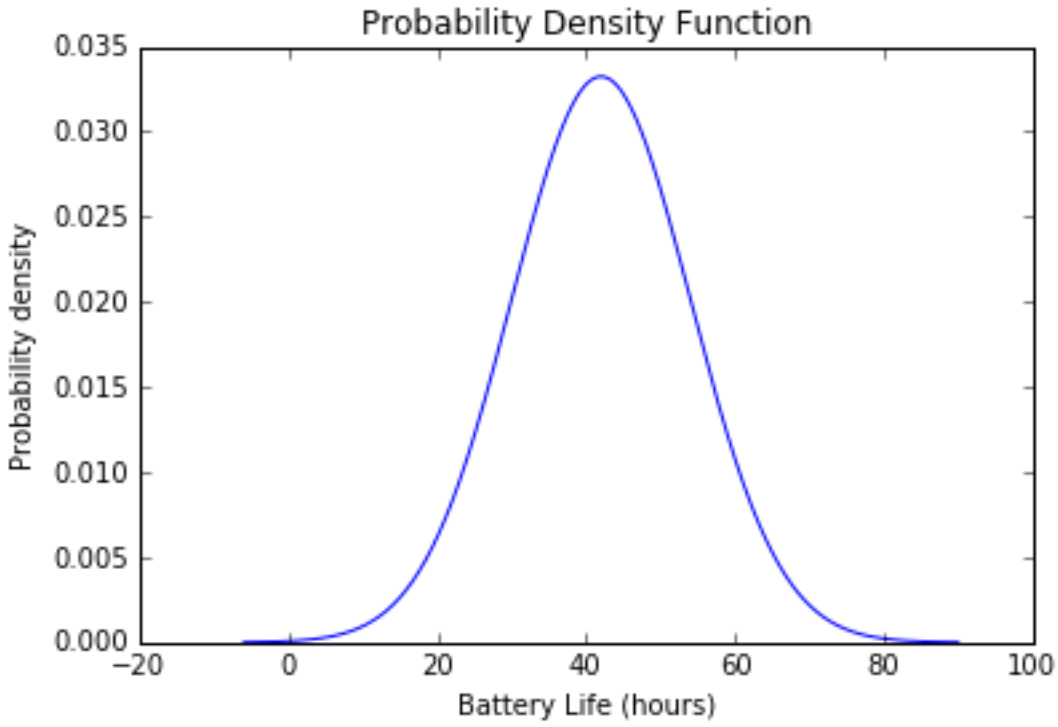
Ω = { BGG, GBG, GGB}

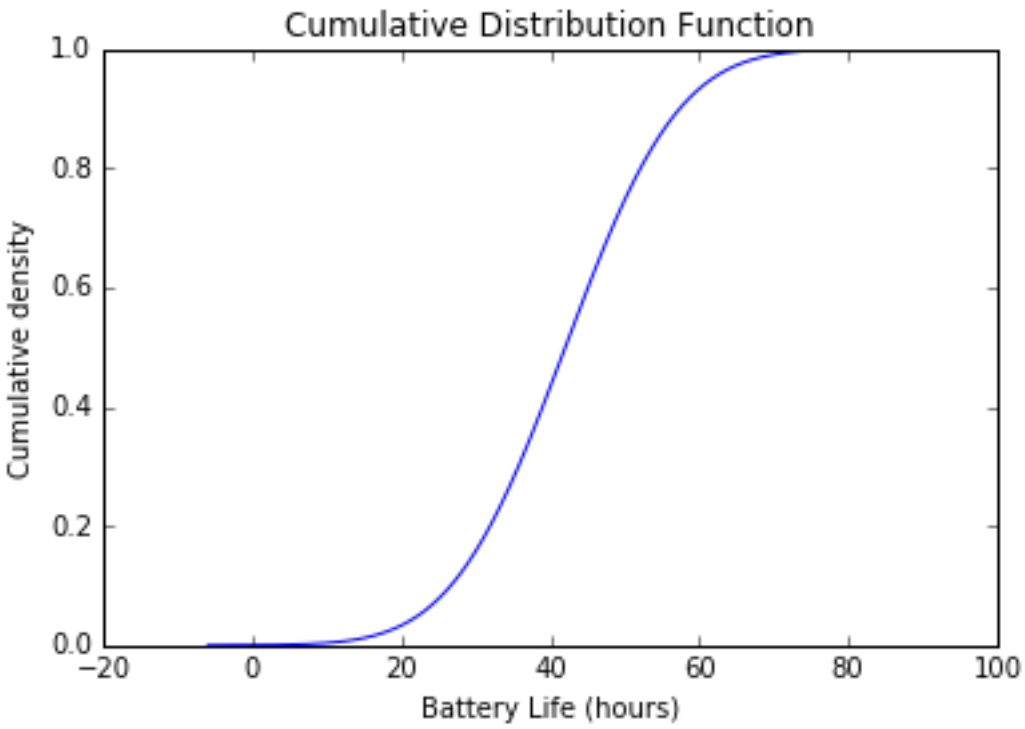
P(child one = B) = 1/3

e. **Annoyingly, your friend won't tell you whether you're right. Instead, she reveals that it's not the middle child. Should you change your answer? Why or why not?**

Monte Halls problem. Initially, P(right) = 1/3 and P(wrong) = 2/3
P(1^{st} child = boy) = 1/3
P(2^{nd} child = boy) = 0 (Revealed)
P(3^{rd} child = boy) = 2/3.
Since, the sets right and wrong don't change, the probability of second child being a boy gets add to the third one.
Hence, better to switch.

3. ***Avoiding the death of your cell-mate.*** **Assume battery-life of an iphone 5 is well-approximated by a normal distribution with $\mu = 42$ hours and $\sigma = 12$ hours (assuming lower-power options disabled).**
   a. Graph the pdf and cdf in python. Paste output below:





Your friend wants to meet in exactly 5 hours, but is not sure where exactly so she will text you 30 minutes before. You last charged your phone 36 hours ago.

b. **Find the probability that your phone will die before your friend txts. (show work).**

Assuming that the phone is still alive after 36 hours,

P(36 < X < 40.5) = $\int_0^{40.5} p(x)dx - \int_0^{36.5} p(x)dx$

Using CDF of the normal distribution in python,

P(X < 40.5) = $ss.norm(mean, SD).cdf(40.5) = 0.4503$

P(X < 36) = $ss.norm(mean, SD).cdf(36) = 0.3085$

Thus, required probability = 0.1418

c. **You think you'll hangout with your friend for 3.5 hours. What is the probability you receive the txt but the phone dies while you are hanging out with your friend? (show work)**

Desired event battery life lies between (36+ 5) 41 hours and (41+3.5) 44.5 hours

P(41< X < 44.5) $= \int_{41}^{44.5} p(x)dx$

$= \int_0^{44.5} p(x)dx - \int_0^{41} p(x)dx$

Using CDF of the normal distribution in python,

P(X < 44.5) = $ss.norm(mean, SD).cdf(44.5) = 0.583$
P(X < 41) = $ss.norm(mean, SD).cdf(41) = 0.467$
Therefore, desired probability = $0.583 - 0.467 = 0.116$

i. **You decide it's too risky to chance your phone dieing before receiving the txt. You insist your friend text you sooner. You want to be 95% confident that your phone is not dead before she texts; In how many hours should you insist she texts by?**
Since, we want to be 95% confident, we need to calculate the the probability that the event occurs with probability of 0.05.

$$P(36 < X < Y = y) = 1 - 0.95$$
$$\Rightarrow \quad P(X < Y = y) - \mathrm{P}(\mathrm{X} < 36) = 0.05$$

$P(X < 36) = ss.norm(mean, SD).cdf(36) = 0.3085$
Therefore, desired probability $P(X < y) = 0.35853753872.$
Using, $y = ss.norm.ppf(0.35853753872)$

$$y = 37.65155233316429$$

Therefore, you insist your friend that she texts you within 1.65 hours so that you're confident that your phone doesn't die before she texts.

4. **Nature's Eye in the Sky.** You are given satellite imagery data describing tree-cover in a particularly den se region of the Amazon. Specifically, the data contain observations for every square kilometer indicating the proportion of tree cover (PTC) (e.g. from 0 = none at all to 1 = fully covered).
   a. **If X is a random variable representing the PTC per square kilometer, should it be continuous or discrete? Why?**
   Continuous. The observations could have infinite values between any two points in [0,1]. Thus, it is modeled using a continuous variable

After plotting the data, you believe it is best best modeled as a random variable, X, with range [0, 1] and probability density function (pdf) $f(x) = Cx^3$.

   b. **What must be the total probability in the range [0, 1]?**

$$\int_0^1 f(x)\,dx = \int_0^1 Cx^3\,dx$$
$$= \lfloor Cx^4/4 \rfloor^1$$
$$= C/4$$

   c. **C is a normalization constant. What must be the value of C in order for this to be a valid pdf with range [0,1]? *(show work)***
   Probability density function must satisfy :
   (i) $f(x) \geq 0$ for all $x \in [0,1]$,
   $\Rightarrow \quad C > 0$
   (ii) $\int_{-inf}^{inf} f(x)\,dx = 1$

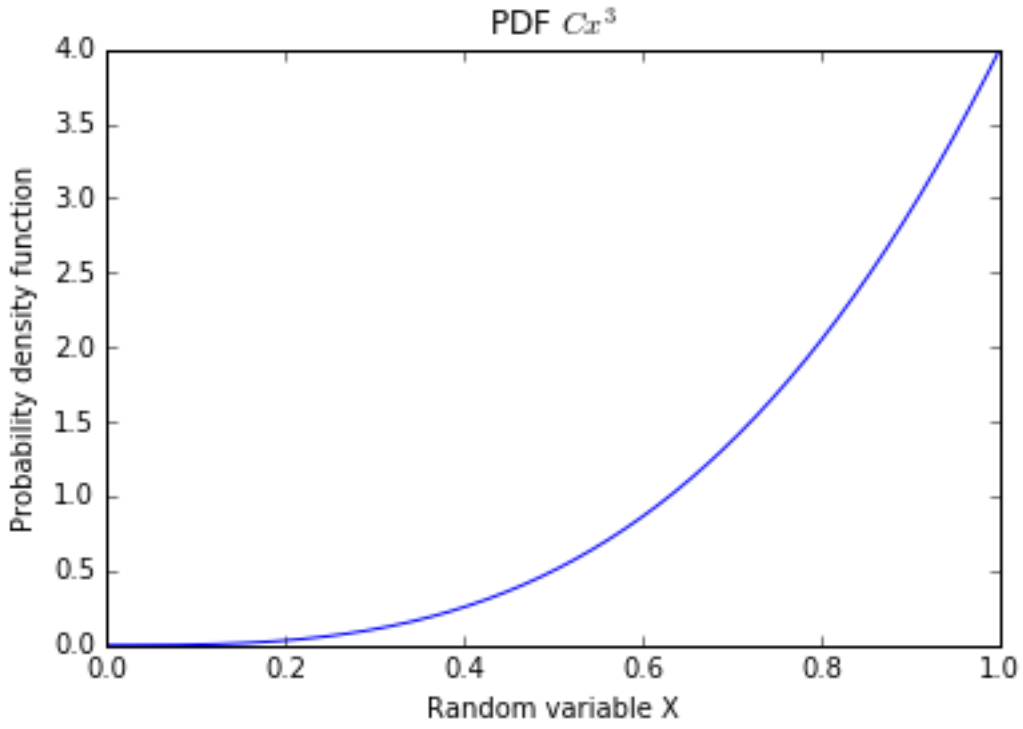
And $f(x) = \begin{cases} Cx^3, & 0 \le x \le 1 \\ 0, & Otherwise \end{cases}$

Therefore,

$$\int_0^1 f(x)\,dx = \int_0^1 Cx^3\,dx$$
$$= \lfloor Cx^4/4 \rfloor^1$$
$$= \lfloor Cx^4/4 \rfloor$$

⇨ *C = 4.*

d. Graph the pdf in python and paste the image (recommend png) below.



The International Center for Rerainforestification (ICR) has determined that any square kilometer with PTC < 0.3 is in "terminal risk". A politician would like to see how growth is coming along and so has gotten out a map and blindly pointed to 4 coordinates he would like to see in person.

e. **Because you (love|hate) the rainforest, you want to know if the politician will see the most damaged areas. According to our model, what is the probability that the politician lands in a square kilometer labeled "terminal risk"?**

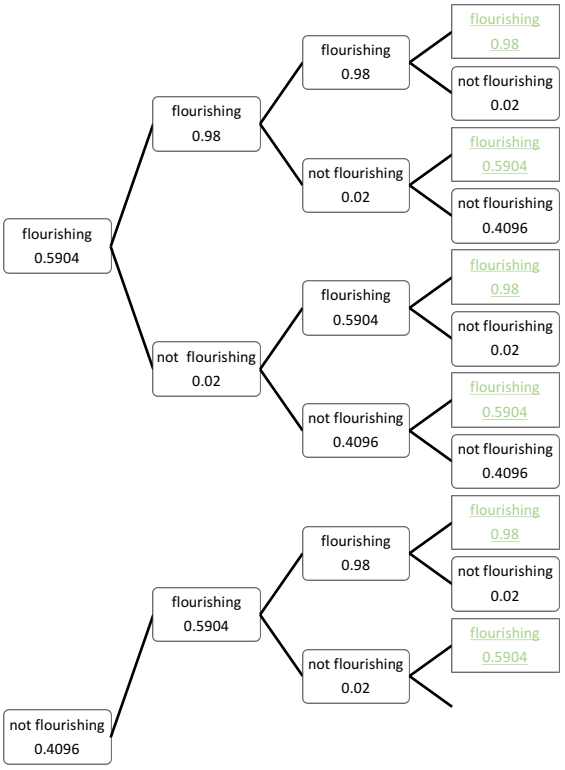
P(PTC < 0.3) = $\int_0^{0.3} f(x)\,dx = \int_0^{0.3} 4x^3\,dx$

$$= x^4|_0^{0.3} = 0.0081$$

Therefore, $P\ (terminal\ risk) = 0.0081$

f. **The ICR labels square kilometers with PTC > .8 as "flourishing". MLE on neighboring pairs of squares indicates that if one is "flourishing" then the other is 98% likely to be flourishing (F ~ Bernoulli (*.98*)). You find out the politician is lazy and just treks straight through 4 contiguous neighboring squares after picking the first square at random. What is the probability that the last square he visits is "flourishing"? State your assumptions and reasons for making them.**

$$P\,(PTC\ >\ 0.8) = 1 - P\,(PTC\ <\ 0.8)$$
$$= 1 - \int_0^1 f(x)\,dx$$
$$= 1 - \int_0^{0.8} 4x^3\,dx$$
$$= 1 - x^4|_{0.8}^1 = 0.5904$$
$$P\,(flourishing)\ =\ 0.5904$$
$$P\,(not - flourishing)\ =\ 0.4096$$
$$P\,(flourishing\,|\,prev =\ flourishing)\ =\ 0.98$$
$$P\,(not - flourishing\,|\,prev =\ flourishing)\ =\ 0.02$$
$$P\,(flourishing\,|\,prev =\ not - \ flourishing)\ = P(flourishing)$$



Therefore, $P\ (S4\ =\ flourishing)$

$= (0.5904)*(0.98)^3 +$
$(0.5904)*(0.98)(0.02)(0.5904) +$
$(0.5904)(0.02)(0.5904)(0.98) +$
$(0.5904)(0.02)(0.4096)(0.5904) +$
$(0.4096)(0.5904)(0.98)^2 +$
$(0.4096)(0.5904)(0.02)(0.5904) +$
$(0.4096)^2(0.5904)(0.98) +$
$(0.4096)^3(0.5904)$
$= 0.5556797568 +$
$0.00682276 +$
$0.00683201433 +$
$0.00342019003 +$
$0.27818009049 +$
$0.00342019003 +$
$0.09707162959 +$
0.06971544123

$= 0.\ 835$

Part II. Programming Assignment

Dataset: 2015 County Health Rankings
http://www.countyhealthrankings.org/sites/default/files/2015%20CHR%20Analytic%20Data.csv
A list of important factors related to health of members of communities.
NOTE: Besides counties, this csv contains state totals. Such rows should be filtered out when working with the data, so that you only work with county data.

Instructions: Complete the steps below. Any numbered items indicate output expected from your code. You should label each output with the same text (e.g. "(1) COLUMN HEADERS:" should appear before listing the columns). Output that is graphical should be given the filenames below, with output indicating it was written to the current directory (e.g. "(4) HISTOGRAM OF POPULATION: wrote a1_4_histpop.png").

A. Read in the CSV.

(1) COLUMN HEADERS: All column headers ending in "Value"
(2) TOTAL COUNTIES IN FILE: The total number of counties in the file. (see "NOTE" above)
(3) TOTAL RANKED COUNTIES: The total number of counties without a "1" in the field "County that was not ranked"

B. Model whether a county was ranked based on its population.

(4) HISTOGRAM OF POPULATION: *a1_4_histpop.png*: A histogram of the field "'2011 population estimate Value'". Choose an appropriate number of bins
(5) HISTOGRAM OF LOG POPULATION: *a1_5_histlog.png*: Add a column, "log_pop" = log("2011 population value"). (Side-note: log transforming the data makes it easier to model.)
(6) KERNEL DENSITY ESTIMATES: *a1_6_KDE.png*: 2 kernel density plots based on log_pop: (a) counties not ranked, and (b) counties ranked. Overlay the density plots over each other into a single graph. Zoom in if necessary to see the the non-ranked distribution clearly.
(7) PROBABILITY RANKED GIVEN POP: Three probabilities --The estimated probability that an unseen county would be ranked, given the following (non-logged) populations: 300, 3100, 5000.

C. Model the health scores as normal. As in (1), consider each column ending in "Value".

(8) LIST MEAN AND STD_DEV PER COLUMN: For each value column, output it's mean and standard deviation according to MLE, assuming a normal distribution (pprint a dictionary of {column: (mean, std-dev), … }). *Hint, MLE tells us:*
$\hat{\mu} = \overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$
(9) PSEUDO-POP-DEPENDENT COLUMNS: List of columns which appear to be dependent on log_pop. We will discuss standard tests for independence after

discussing hypothesis testing. For the purposes of this assignment, we will call two continuous random variables, A and B "pseodo-independent" iff $|\mathrm{E}(A|\ B<\mu_B) - \mathrm{E}(A|\ B>\mu_B)| < 0.5 * {}_A$. Although the variables in "value" columns have been normalized by population, some may still be dependent on population. For example, certain mortality rates are higher in more rural communities because they are often poorer and have less access to health care.

Python File: Your python 2.7 script should be called "a1_STUDENTID.py", and it should run with the command "python a1_STUDENTID.py FILE.csv". It should be self-contained (use classes within to keep organized if you like). Your code should be well documented with comments and not import any mathematical, statistical, and data scientific libraries beyond: numpy, scipy, pandas, matplotlib, and the default libraries that come with Python (ask if you are not sure). *If you are developing in Python Notebook or Jupyter, make sure to export your source code as a .py file and test that it runs in the Python 2.7 compiler (i.e. not iPython) before submitting.*

Testing: Code will be tested against the csv listed above as well as a version with random modifications to the data, including replacement of values with new values and removal / insertion of rows. Please be sure your code can handle such modifications to the csv.

Submission: Your single python file as well as a pdf containing your solutions to Part I, should be submitted via Blackboard by the due date listed at the top.