Homework_12

December 4, 2019

```
In [1]: # HIDDEN
        from datascience import *
        from prob140 import *
        import numpy as np
        import matplotlib.pyplot as plt
        plt.style.use('fivethirtyeight')
        %matplotlib inline
        from scipy import stats
        from scipy import special
        import warnings
        warnings.filterwarnings('ignore')
```

1 Homework 12

1.0.1 Instructions

Your homeworks have two components: a written portion and a portion that also involves code. Written work should be completed on paper, and coding questions should be done in the notebook. You are welcome to LaTeX your answers to the written portions, but staff will not be able to assist you with LaTeX related issues. It is your responsibility to ensure that both components of the homework are submitted completely and properly to Gradescope. Refer to the bottom of the notebook for submission instructions.

1.0.2 Preliminary: Distributions

The code below will help you in some of the exercises. Please refer to it as needed.

Discrete Distribution Table Defined by a Probability Function

`Table().values(values_array).probability_function(name_of_function)` evaluates to the distribution table of a discrete random variable X such that:

- `values_array` is an array consisting of the possible values of X
- `name_of_function` is the name of a function that takes x as its argument and returns $P(X = x)$

Beta Density

For an array `x` of values in the unit interval, `stats.beta.pdf(x, r, s)` evaluates to the array f(x) where f is the beta (`r`, `s`) density.

1.0.3 Preliminary: Line Plots

In Data 8 you used `Table.plot` to draw line plots, but in this class the line plots have typically been drawn using the `plot` function of `matplotlib`. Here is the syntax; you are going to need it in the exercises. It's easier than setting up tables first, especially when you want to overlay multiple plots.

The `pyplot` module of `matplotlib` has been imported for you as `plt`. This is its standard abbreviation.

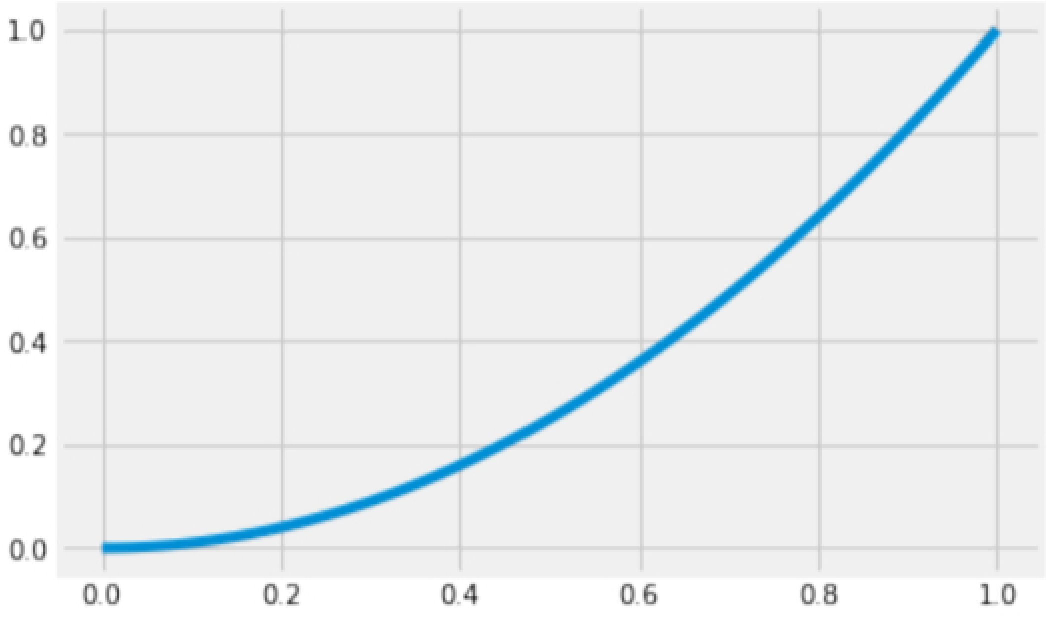
Let `x` and `y` be two numerical arrays of the same length. Then `plt.plot(x, y)` produces a line plot with values of `x` on the horizontal axis and the corresponding values of `y` on the vertical. The plot "joins the dots" (`x.item(0)`, `y.item(0)`), (`x.item(1)`, `y.item(1)`), (`x.item(2)`, `y.item(2)`), and so on.

The optional argument `lw=n` can be used to set a line width of `n` units. Just use `lw=2` in this homework.

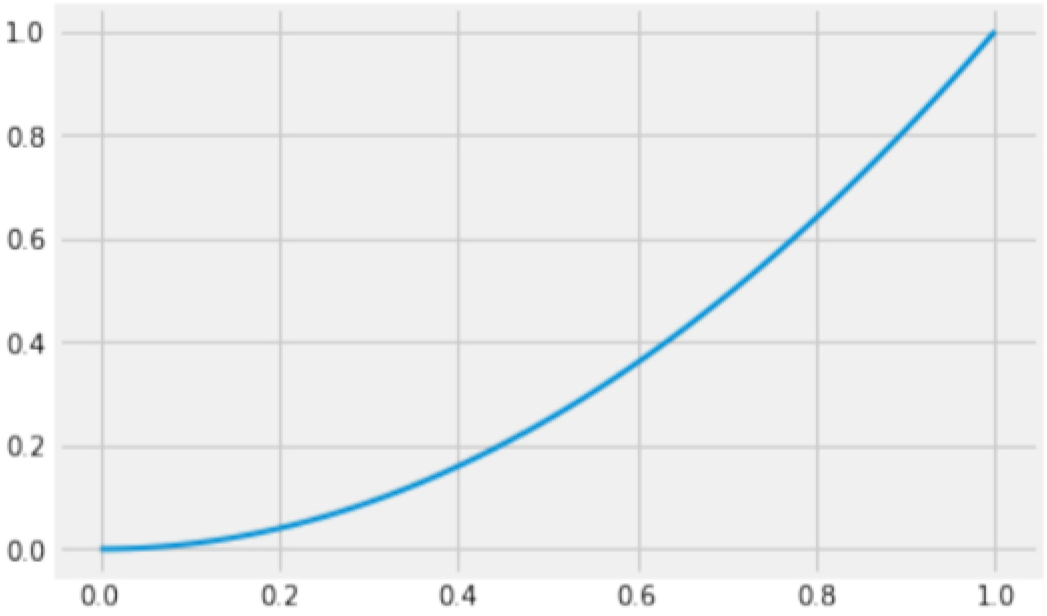
The semicolon at the end of the last line of code in each cell stops `matplotlib` from blurting out text that we don't need here.

Run these cells to see some examples. Note the overlaid plots: they are straightforward to draw, and Python chooses a different color for each plot.

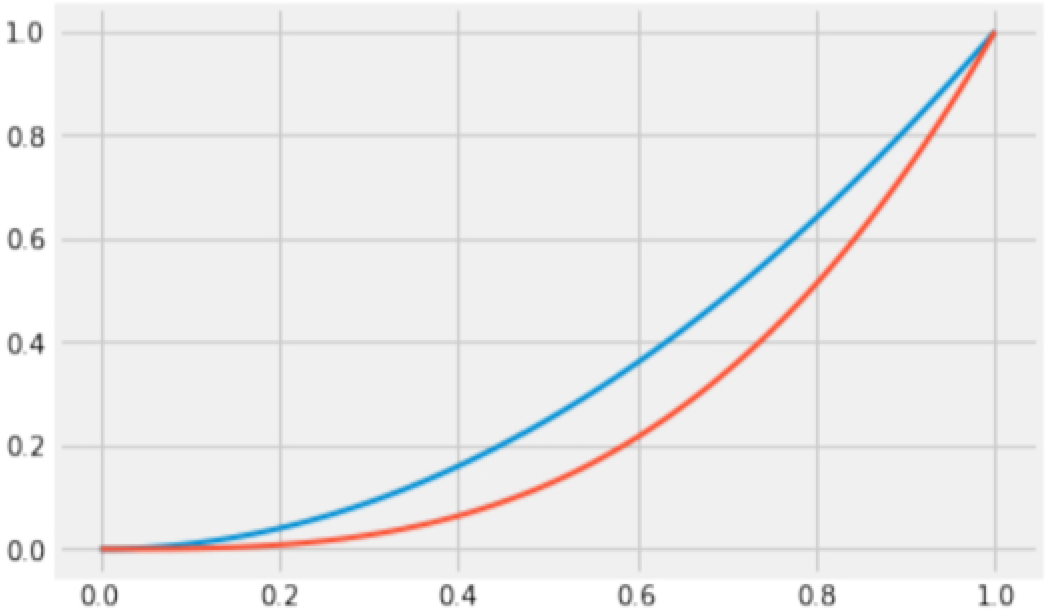
```
In [2]: x = np.arange(0, 1.01, 0.01)
        x_squared = x ** 2
        plt.plot(x, x_squared);
```



```
In [4]: plt.plot(x, x_squared, lw=2);
```

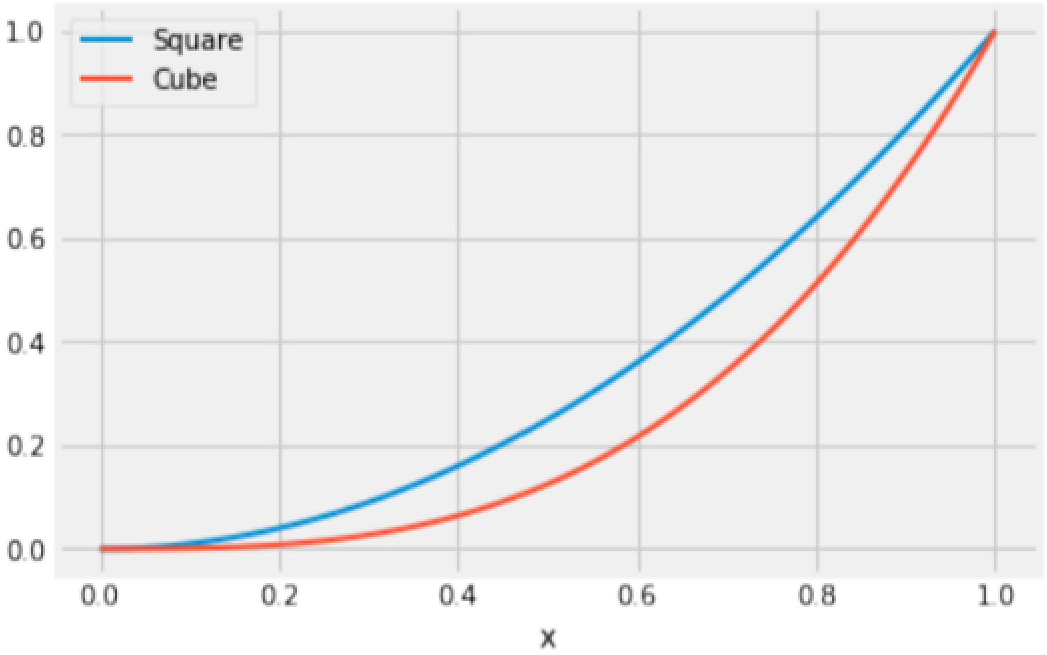


```
In [5]: x_cubed = x ** 3
        plt.plot(x, x_squared, lw=2)
        plt.plot(x, x_cubed, lw=2);
```



You can get more fancy with labels and legends, but you don't have to do that in this homework.

```
In [6]: x_cubed = x ** 3
        plt.plot(x, x_squared, lw=2, label = 'Square')
        plt.plot(x, x_cubed, lw=2, label = 'Cube')
        plt.legend()
        plt.xlabel('x');
```



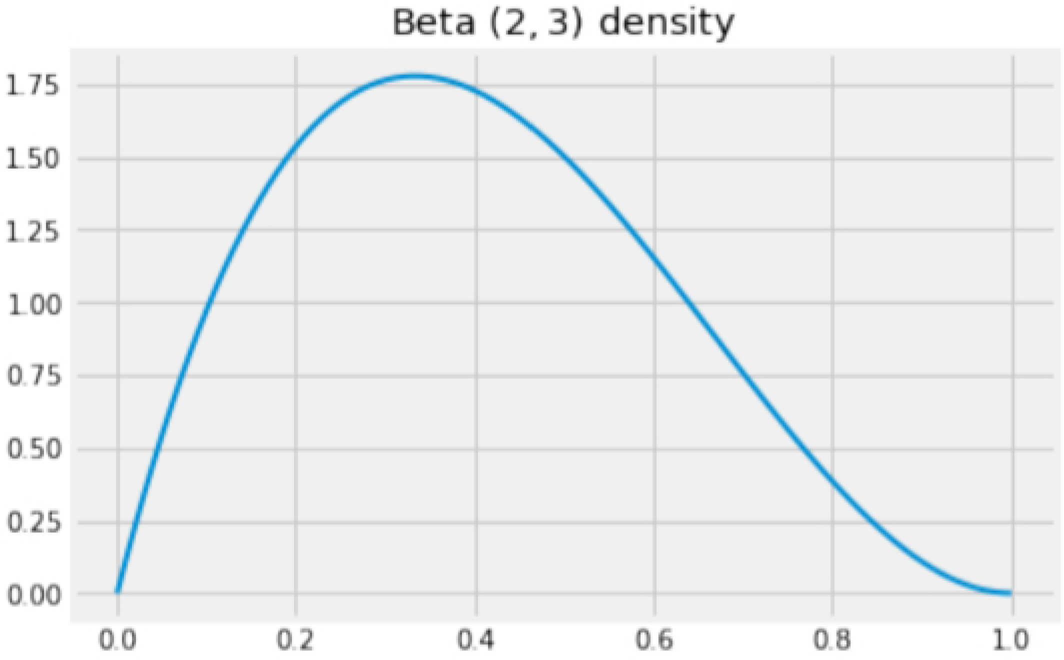
This is only a tiny fraction of what you can do with `matplotlib`. For more, here is a great tutorial by Nicolas Rougier.

1.0.4 1. Heads in Tosses of a Random Coin

Let X be a random proportion with a prior distribution that is beta $(2, 3)$. Given $X = p$, let N be the number of heads in 5 tosses of a p-coin.

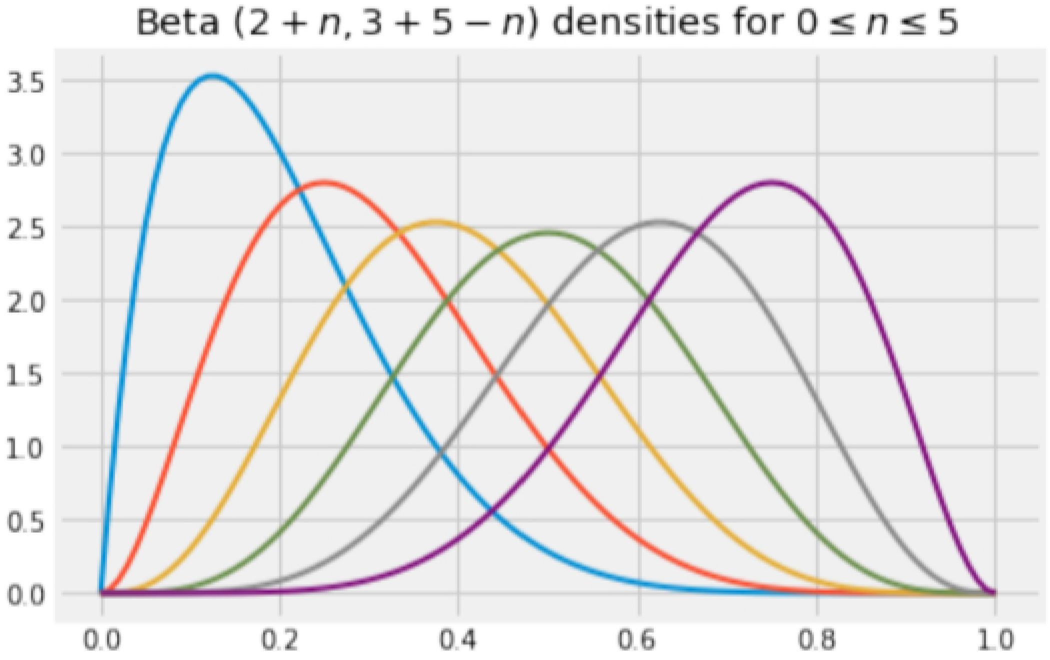
a) Plot the prior density of X.

```
In [6]: # 1a


        x = np.arange(0, 1.01, 0.01)
        plt.plot(x, stats.beta.pdf(x, 2, 3), lw=2)
        plt.title('Beta $(2, 3)$ density');
```



b) For each $n = 0, 1, 2, 3, 4, 5$, plot the posterior density of X given $N = n$. All six plots should be on the same graph.

```
In [7]: # 1b

        x = np.arange(0, 1.01, 0.01)
        for n in np.arange(0, 6, 1):
            plt.plot(x, stats.beta.pdf(x, 2+n, 3+5-n), lw=2)
        plt.title('Beta $(2+n, 3+5-n)$ densities for $0 \leq n \leq 5$');
```



c) For each $n = 0, 1, 2, 3, 4, 5$, find the MAP estimate of the probability of heads given $N = n$, by completing the cell below with an appropriate expression. Here `map_estimates` is an array that contains all the MAP estimates.

By eye, check that your answers are consistent with the graphs in (b).

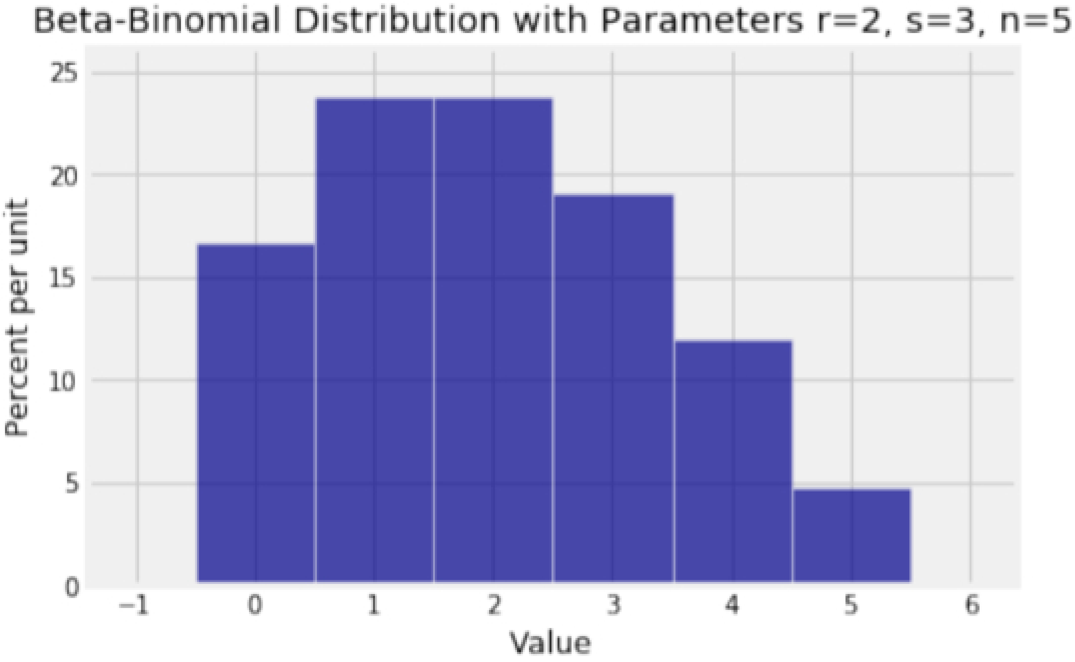
```
In [8]: # 1c


        r = 2
        s = 3
        n = 5
        k = np.arange(0, n+1)

        map_estimates = (r + k - 1) / (r + s + 5 - 2)


        map_estimates

Out[8]: array([0.125, 0.25 , 0.375, 0.5  , 0.625, 0.75 ])
```

d) Complete the cell below to display the probability histogram of N. You can use any number of lines of code in the definition. In the blanks in the last line of the cell, enter the name of the distribution and provide the numerical parameters.

If you want to, you can use `special.factorial(j)` to get the factorials of the non-negative integer array `j`. The `special` module of `SciPy` has been imported for you. But you can also write the code without computing factorials.

```
In [9]: # 1d
```

```
def probs_N(n):
    prob = (n+1)*(7-n)*(6-n)*(1/(9*28))
    return prob

dist = Table().values(np.arange(0, 6, 1)).probability_function(probs_N)
Plot(dist)
plt.title('Beta-Binomial Distribution with Parameters r=2, s=3, n=5');
```



e) Find the numerical value of $E(N)$ without using a code cell.
1e) Given $X = x$, the conditional distribution of N is binomial (n, x). So $E(N|X) = 5X$.
$E(N) = E(5X) = 5 \cdot \frac{2}{2+3} = 2$
#newpage

1.0.5 2. Two-Colored Die

A die that has two green faces and four blue faces is rolled repeatedly. Let R be the number of rolls till both colors have appeared.
a) Find $E(R)$.
b) Find $SD(R)$.

1.0.6 [Solution] Two-Colored Die

Condition on the first roll. Formally, condition on I which is the indicator of the first roll being blue.

- If $I = 0$ then $R = 1 + W_b$ where W_b is the additional number of rolls till a blue face appears.

- If $I = 1$ then $R = 1 + W_g$ where W_g is the additional number of rolls till a green face appears.

Here W_b and W_g are independent but not identically distributed geometric random variables on $\{1, 2, \ldots\}$.

By Section 22.3, if X has geometric (p) distribution on $\{1, 2, \ldots\}$ then $E(X) = 1/p$ and $Var(X) = q/p^2$. So:

- $E(R \mid I = 0) = 1 + 3/2 = 5/2$ and $Var(R \mid I = 0) = (1/3)/(4/9) = 3/4$
- $E(R \mid I = 1) = 1 + 3/1 = 4$ and $Var(R \mid I = 1) = (2/3)/(1/9) = 6$

5a) $E(R) = (1/3)(5/2) + (2/3)(4) = 3.5$.
5b) $Var(R) = \big[(1/3)(3/4) + (2/3)(6)\big] + \big[(1/3)(25/4) + (2/3)(16) - 3.5^2\big] = 4.75$.
So $SD(R) \approx 2.18$.
#newpage

1.0.7 3. Chinese Restaurant Process: Long Run Proportion at Table 1

The goal of this exercise is for you to explain a density that you observed at the end of Lab 10.

Parts **a** and **b** are about the "beta binomial" process we studied extensively in class. You can just apply the results derived in class. You'll find them in Chapter 21 of the textbook.

a) Define the "Beta $(1, \theta)$ Binomial" process as follows:

- X has beta $(1, \theta)$ distribution.
- Given $X = p$, there is a sequence of i.i.d. Bernoulli (p) trials $I_1, I_2, …$.
- $S_n = I_1 + I_2 + \cdots + I_n$ is the number of successes in the first n trials.

Fill in the blank with the name of a distribution and the appropriate parameters:

As n gets large, with probability 1 the proportion $\frac{S_n}{n}$ approaches a limit. The distribution of this limit is ________________.

b) For the Beta $(1, \theta)$ Binomial process, find the transition probability $P(S_n = k + 1 \mid S_{n-1} = k)$. Given $S_{n-1} = k$, what are the other possible values of S_n and what are their probabilities?

c) Now refer to Lab 10 for the description of Chinese Restaurant process with parameter θ. As in the lab, let W_n be the number of people at Table 1 at time n. The goal is of this part of the exercise is to figure out the transition behavior of W_n.

Because Table 1 starts deterministically with Person 1 at time 1, we have to be a bit careful about what's random and what's constant. Let's consider $V_n = W_n - 1$, the random number of people *other than Person 1* who are at Table 1 at time n.

Find $P(V_{n+1} = k + 1 \mid V_n = k)$. This is a straightforward application of the rules of the Chinese Restaurant process, but you have to be careful about how many people are at Table 1 when Person $n + 1$ enters the system.

d) Given $V_n = k$, what are the possible values of V_{n+1}? Compare the transition behavior of the following two sequences:

- $V_2, V_3, …$ of the Chinese Restaurant process with parameter θ
- $S_1, S_2, …$ of the Beta $(1, \theta)$ Binomial process

Then explain why the distribution of the long run proportion of people at Table 1 of the Chinese Restaurant process is beta $(1, \theta)$ where θ is the parameter of the process. That is, recall that you saw in the lab that the paths of $\frac{W_n}{n}$ converge as $n \to \infty$, and explain why the limit of $\frac{W_n}{n}$ as $n \to \infty$ has the beta $(1, \theta)$ distribution.

1.0.8 [Solution] Chinese Restaurant Process

a) beta $(1, \theta)$.

b) Apply the formula derived in Section 21.1 to get $P(S_n = k+1 \mid S_{n-1} = k) = \frac{1+k}{1+\theta+(n-1)} = \frac{1+k}{n+\theta}$.

Given $S_{n-1} = k$, the only other possible value of S_n besides $k+1$ is k. So $P(S_n = k \mid S_{n-1} = k) = 1 - \frac{1+k}{n+\theta}$.

c) There are n people in the system at time n. Given $V_n = k$, there are $1+k$ people at Table 1 at time n. By the rules of the process, the chance that Person $n+1$ picks Table 1 is proportional to $1+k$ and hence is equal to $\frac{1+k}{n+\theta}$.

d) Given $V_n = k$, the possible values of V_{n+1} are $k+1$ or k. The transition behavior is exactly the same as that of the S_{n-1} sequence of the Beta $(1, \theta)$ Binomial process.

When n is large, $\frac{W_n}{n} \approx \frac{V_n}{n}$ because W_n and V_n are off by the constant 1.

Also when n is large, $\frac{V_n}{n}$ of the CR process has the same behavior as $\frac{S_n}{n}$ of the Beta $(1, \theta)$ Binomial process; the time index being off by 1 doesn't matter in the limit.

In the Beta-Binomial process, the proportions $\frac{S_n}{n}$ converge to the beta $(1, \theta)$ distribution as n gets large.

So the limit of $\frac{W_n}{n}$ as n gets large must have the beta $(1, \theta)$ distribution.

#newpage

1.0.9 4. Sections

A class of 60 students has three sections. Summary statistics for scores on Quiz 7: - Section 1: 25 students, mean 23, SD 3 - Section 2: 20 students, mean 25, SD 2 - Section 3: 15 students, mean 27, SD 4

Let S be the Quiz 7 score of a student picked at random from the class. Use the code cell below to calculate numerical answers for parts a and b.

a) [WRITTEN + CODE] Find $E(S)$.

b) [WRITTEN + CODE] Find $SD(S)$.

1.0.10 [Solution] Sections

a) $E(S) = (25/60) * 23 + (20/60) * 25 + (15/60) * 27$

$E(S) = 24.6666$

b) $Var(S) = E(S^2) - (E(S))^2$

$= (25/60)3^2 + (20/60)2^2 + (15/60)4^2 + (25/60)23^2 + (20/60)25^2 + (15/60)27^2 - E(S)^2$

$SD(S) = \sqrt{Var(S)}$

$SD(S) = 3.412$

```
In [7]: #4a

        E_s = (25/60)*23 + (20/60)*25 + (15/60)*27
        E_s

Out[7]: 24.666666666666664

In [8]: #4b


        var_s = (25/60)*3*3 + (20/60)*2*2 + (15/60)*4*4 + (
    25/60)*23*23 + (20/60)*25*25 + (15/60)*27*27 - E_s*E_s
sd_s = np.sqrt(var_s)
sd_s
```

```
Out[8]: 3.4115815817431403
```

#newpage

1.0.11 5. Uniform Stick-Breaking

A point X is chosen uniformly from the interval $(0, 10)$ and then a point Y is chosen uniformly from the interval $(0, X)$. This can be imagined as snapping a stick of length 10 and then snapping one of the broken bits. Stick-breaking processes are related to the Chinese Restaurant process.

a) Find $E(X)$ and $Var(X)$. See Section 15.3 of the textbook for the variance of the uniform.

b) Find $E(Y)$ and $Var(Y)$ by conditioning on X.

1.0.12 [Solution] Uniform Stick Breaking

a) $E(X) = 5$

$Var(X) = \frac{(10-0)^2}{12} = \frac{100}{12}$

b) $E(Y) = E(E(Y \mid X)) = E(\frac{1}{2}X) = \frac{5}{2}$

$Var(Y) = E(Var(Y \mid X)) + Var(E(Y \mid X)) = E(\frac{X^2}{12}) + Var(\frac{1}{2}X) = \frac{1}{12}(\frac{100}{12} + 5^2) + (\frac{1}{4})(\frac{100}{12})$

1.1 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

1.1.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using an application. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.

1.1.2 Code Portion

- Save your notebook using File > Save and Checkpoint.
- Generate a PDF file using File > Download as > PDF via LaTeX. This might take a few seconds and will automatically download a PDF version of this notebook.
    - If you have issues, please make a follow-up post on the general HW 12 Piazza thread.

1.1.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 12 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

1.1.4 We will not grade assignments which do not have pages selected for each question.