Homework 1

Name: Kalpita Dapkekar

Student ID : 001446977

1. Data Analysis

The number of students admitted to college A as freshmen is 1000 and on average 15% of the students drop every year of college (assume all degrees take 4 years and that 15% of students in Y1 do not continue to Y2, and then 15% of students in Y2 similarly do not continue to Y3 and so on).

a. **Plot the Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) of the random variable corresponding to year in college of a student (Hint: X ∈ {1, 2, 3, 4}).**

Solution:

Y1 = 1000

PMF : f(1) = $\frac{1000}{3186}$ = 0.313

$15\%\ of\ Y1$ = 150   Y2 = 1000 – 150 = 850

f(2) = $\frac{850}{3186}$ = 0.266

$15\%\ of\ Y3$ = 127

Y3 = 850 – 127 = 723

f(3) = $\frac{Y3}{3186}$ = $\frac{723}{3186}$ = 0.226

15% of Y3 = 108 Y4 = 723 – 108 = 615

f(4) = $\frac{Y4}{3186}$ = $\frac{615}{3186}$ = 0.192

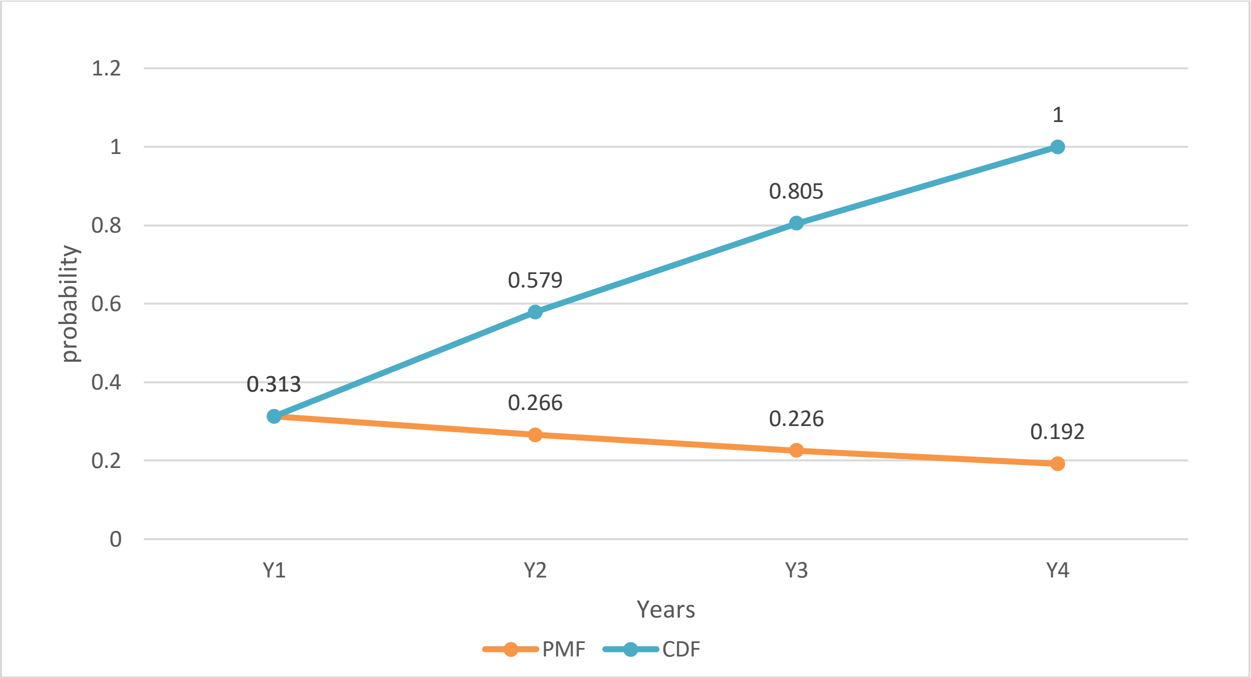
CDF:

F(1) = P(X≤1) = f(1) = 0.313

F(2) = P(X≤2) = f(1)+f(2) = 0.313 + 0.266 = 0.579

F(3) = P(X≤3) = f(1)+f(2)+f(3) = 0.313 + 0.266 + 0.226 = 0.805

$F(4) = P(X \leq 4) = f(1)+f(2)+f(3)+f(4)= = 0.313 + 0.266 + 0.226 + 0.193= 1$



b. **Calculate the expected year in college of a student (mean) and its variance. Show the steps of your calculation.**

Solutions:

$$
\begin{aligned}
E(X) = \mu = \sum xf(x) &= 1f(1) + 2f(2) + 3f(3) + 4f(4)\\
&= 1(0.313) + 2(0.266) + 3(0.226) + 4(0.192)\\
&= 2.291
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2 = E\,[(x-\mu)^2] = \sum(x-\mu)^2 f(x) &= (1-2.291)^2f(1) + (2-2.291)^2f(2) + (3-2.291)^2f(3)\\
&\quad + (4-2.291)^2f(4)\\
&= 1.66(0.313) + 0.8281(0.266) + 0.50(0.226) + 2.92(0.192)\\
&= 1.41
\end{aligned}
$$

c. **Suppose that instead of 15%, α% of students drop out per year (0 ≤ α ≤ 100). What is the mean as a function of α.**

Solutions:

Y1 = 1000

Y2 = $1000 - 10\propto$

Y3 = $\frac{\left(10^5 - 2*1000\propto+10(\propto)^2\right)}{100}$

Y4 = = $\frac{\left(10^7 - 3*10^5\propto+3*10^3(\propto)^2+10(\propto)^3\right)}{10^4}$

E(x) = μ = 1f(1) + 2f(2) + 3f(3) + 4f(4)

$$= \frac{\left(10^7 - 2*10^5\propto+15*10^2(\propto)^2-4(\propto)^3\right)}{4*10^6-6*10^4\propto+4*10^2(\propto)^2-(\propto)^3}$$

d. **Suppose college A merges with nearby medschool B to create university C. In the year of the merge, medschool B has 40 students in year 6, 50 students in year 7, and 100 students in year 8.** Calculate the new expected value of year in college at the combined university C. Is the mean (as compared to part b) stable or sensitive to these new data points? What other statistical measures are there to estimate the average behavior? Are they less or more stable with regards to the outliers introduced in the merger? Justify your answer by computing those alternative measures for the original college A and the merged university C.

Solutions:

f(6) = $\frac{40}{190}$ = 0.210

f(7) = $\frac{50}{190}$ = 0.263

f(8) = $\frac{100}{190}$ = 0.526

E(Y) = μ = 6f(6) + 7f(7) + 8f(8)
= 6(0.210) + 7(0.263) + 8(0.526)
= 1.26 + 1.841 + 4.208
= 7.309

Therefore, The expected value of the sum of several random variables is equal to the sum of their expectations, e.g.,

$$E[X+Y] = E[X]+ E[Y]$$
$$= 2.291 + 7.309 = 9.6$$

The new Expected value of **University C** is 10.497 and of A is 2.291. That means, **mean** value is not stable to these new data points.

Other statistical measure is **Variance.**

$\sigma^2 = E[(y-\mu)^2] = \sum(y-\mu)^2 f(y) = (6-7.309)^2(0.210) + (7-7.309)^2(0.263) + (8-7.309)^2(0.526)$
$= 0.359 + 0.0251 + 0.251$
$= 0.6351$

$$\sigma^2 = \sigma^2_x + \sigma^2_y$$
$$=1.3047 + 0.6351$$
$$= 1.9398.$$

Variance for university C is 1.9398 and for A its 1.41. **Variance** is also Stable to these new data points.

Median:

Median for university A:

A = {615,723,850,100} Median = $\frac{850+723}{2}$ = 786

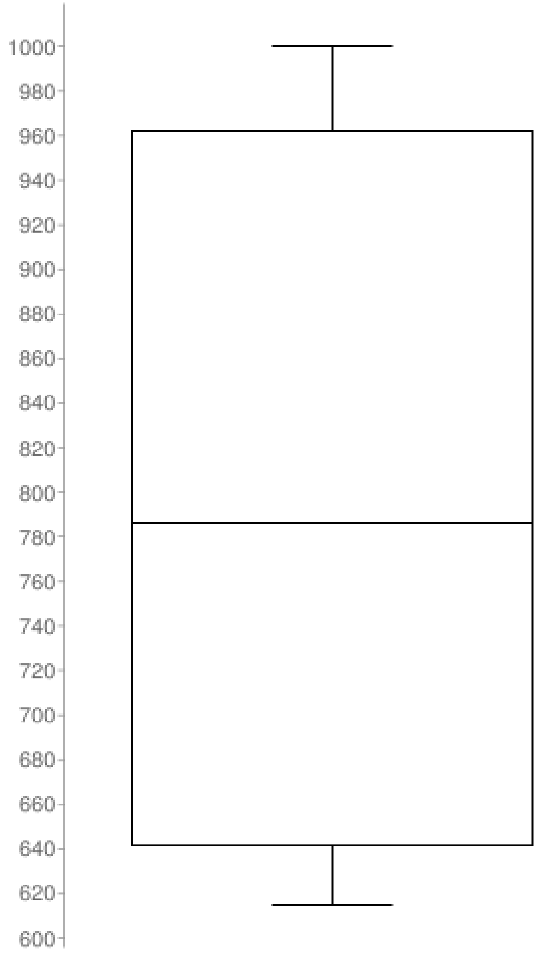
Median for university C:

C = {40,50,100,615,723,850,100} Median = 614

Median is not Stable.

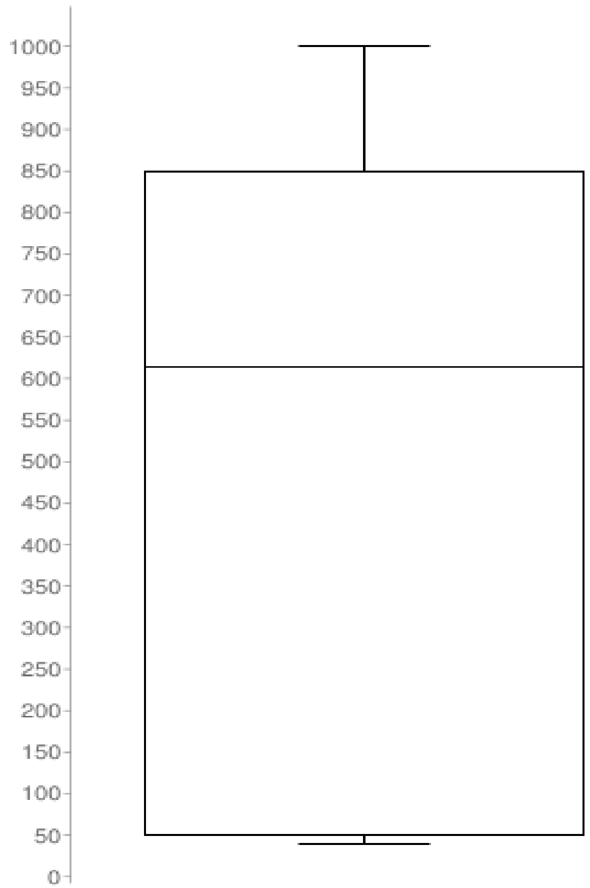
e. **Create a box plots for the years in college variables X_A for college A and X_C university C. (don't use python libraries for this - just draw it by hand or in your submission). Look here: http://www.physics.csbsju.edu/ stats/box2.html for examples of box plots. You should have two boxes one for X_A and one for X_C with their specific statistics.**

Solutions:



Box plot for University A

Sample size: 4
Median: 786.5
Minimum: 615
Maximum: 1000
First quartile: 642
Third quartile: 962.5
Interquartile Range: 320.5
Outliers: none



Box plot for University C:

Sample size: 7
Median: 615
Minimum: 40
Maximum: 1000
First quartile: 50
Third quartile: 850
Interquartile Range: 800
Outliers: none

2. Irreducible data example

a. Discuss the cases when PCA will fail.

Solutions:

i. **Case i:** PCA requires Linear data to reduce dimensionality. If the data is not linear it fails to maximize the variance and to reduce dimensionality. We require non linear data as well to reduce dimensionality. Standard PCA will fail for non linear data. Kernel PCA overcomes the limitations of Standard PCA.

ii. **Case ii:** In the definition of PCA itself it is mentioned that '*PCA find orthogonal PCA components*'. Sometimes data need non orthogonal PCA components. PCA fails to find non orthogonal principal components. Non orthogonal components are called Independent components. This limitation is overcome by Independent component analysis.

b. How do we quantify that it fails?

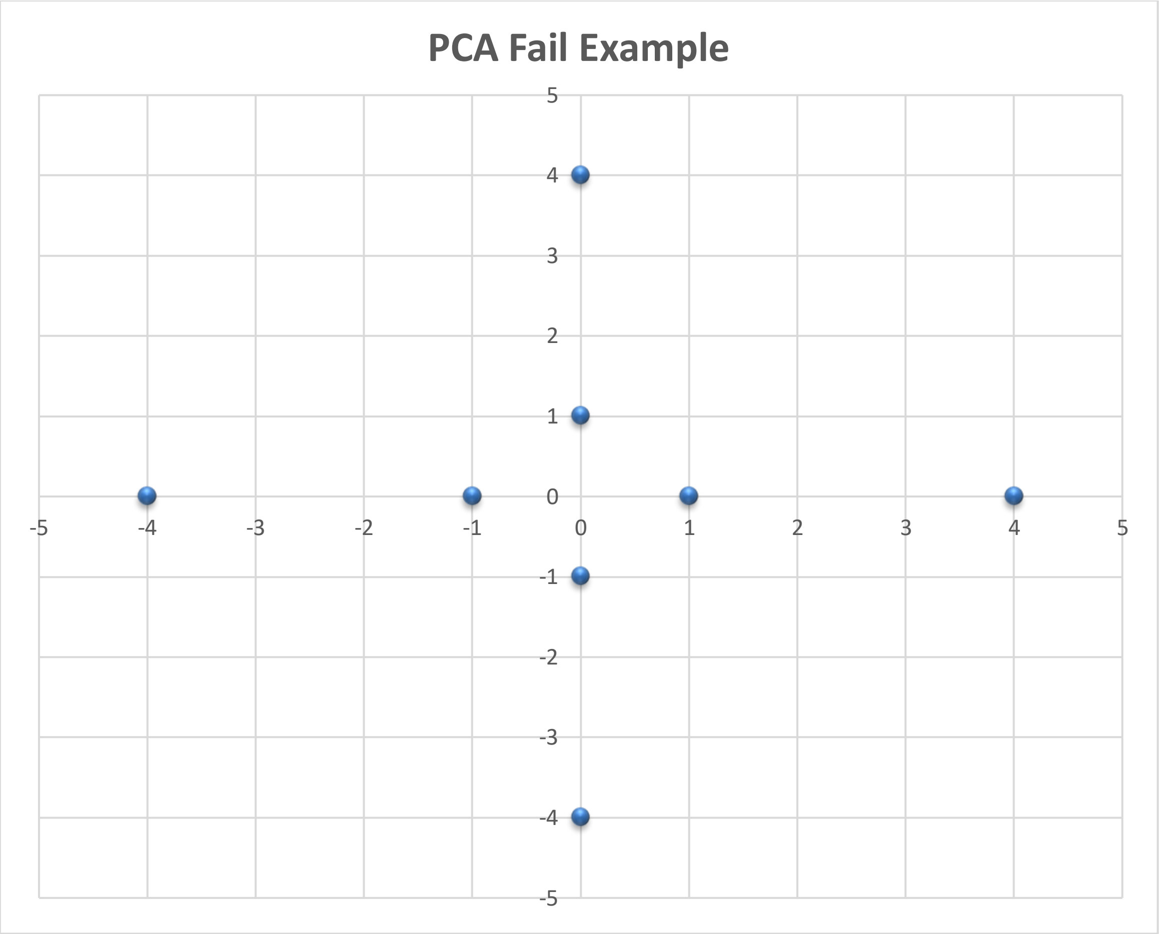
Solutions:

If the data is in different direction and PCA is in different direction then we can say that it failed.

c. Provide a minimal example of a dataset (specify the points as vectors of numbers) in which PCA will not work well for dimensionality reduction. Explain why. Hint: Think of 2D points and reduction to 1D.

Solutions:

Let's consider the following dataset:
(1,0),(0,-1),(-1,0),(0,1) & (4,0),(0,-4),(-4,0),(0,4)



PCA Fail Example

3. Dimensionality reduction

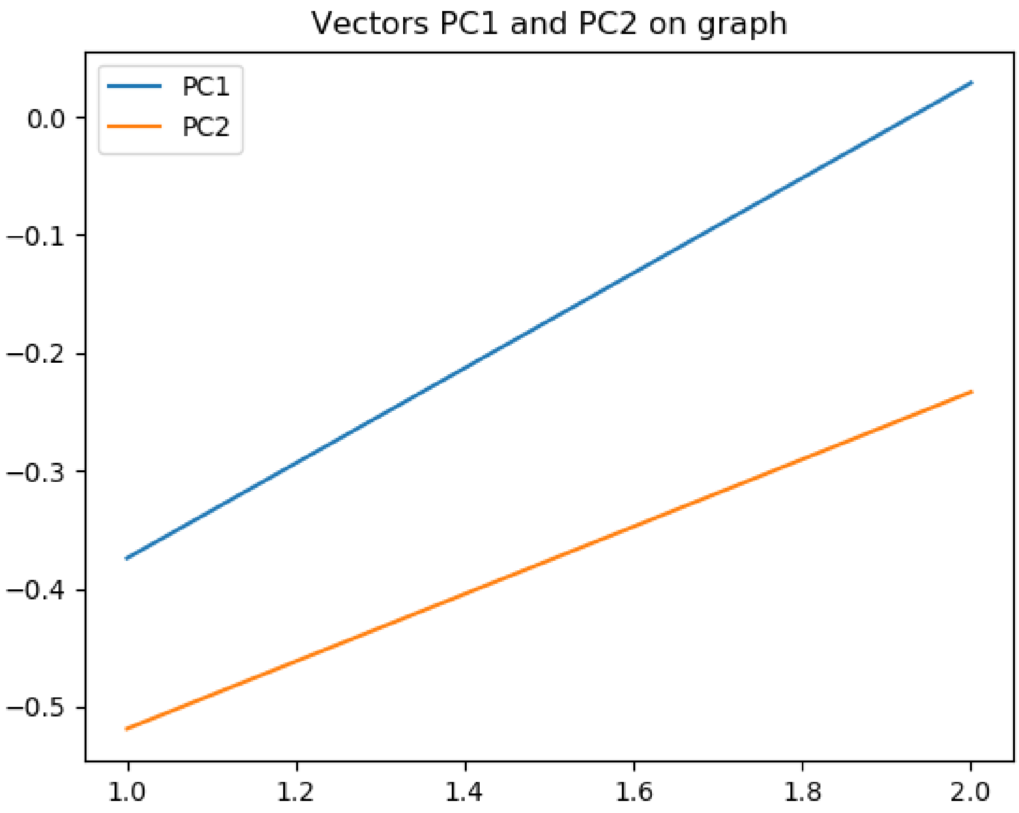
(b)

0.234000205994 seconds for Covariance matrix using numpy library.

0.0103456789656 seconds for Covariance matrix using inner product.

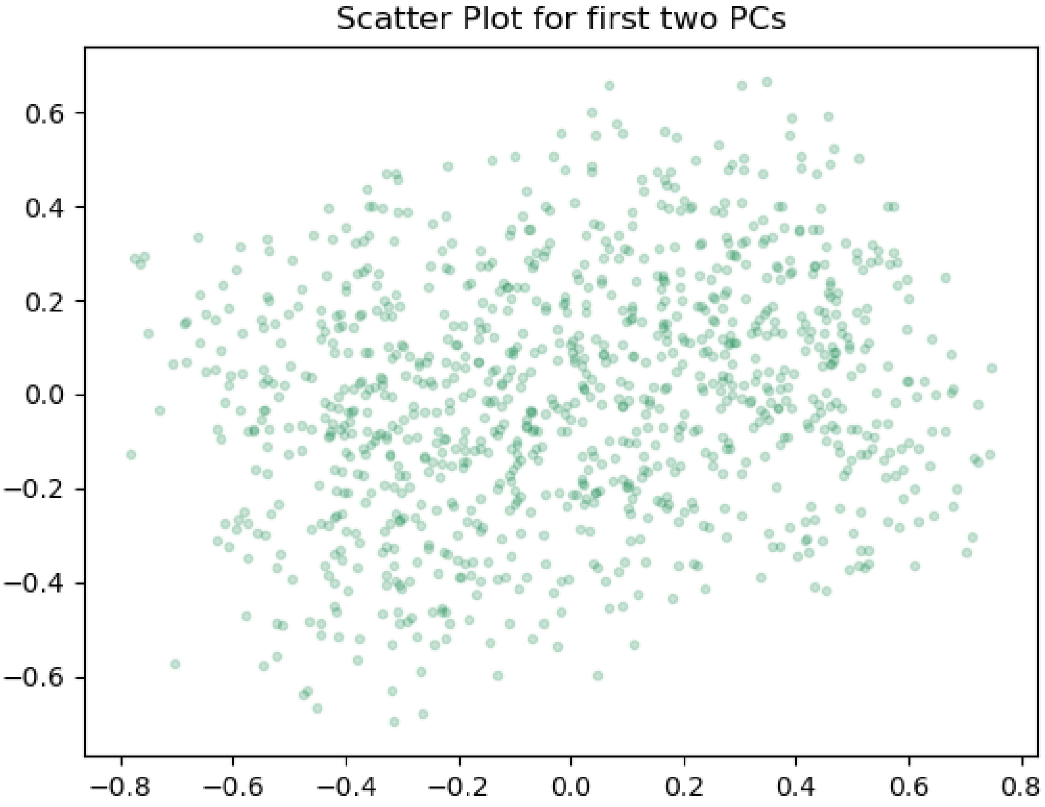
0.0309998989105 seconds for Covariance matrix using outer product.

(e) Line plot of first two components:



(f) Scatter Plot:

As you can see in the following scatter plot clusters are not formed. Data is distributed ununiformly all over the plot.


Scatter Plot for first two PCs
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8