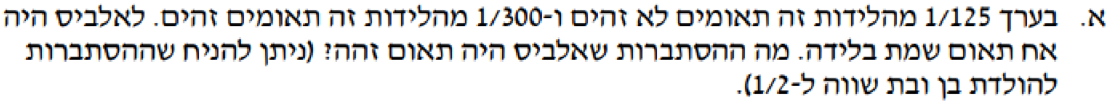
Introduction to data science final project

Part 1

Probability and Bayes' theorem

Question 1

Part 1



א. בערך 1/125 מהלידות זה תאומים לא זהים ו-1/300 מהלידות זה תאומים זהים. לאלביס היה
אח תאום שמת בלידה. מה ההסתברות שאלביס היה תאום זהה? (ניתן להניח שההסתברות
להולדת בן ובת שווה ל-1/2).

We are looking for P(identical twins/ twins)=?

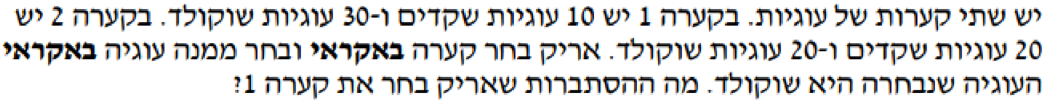
$$P(Frateranl\ twins) = \frac{1}{125}$$

$$P(Identical\ twins) = \frac{1}{300}$$

$$P(twins) = \frac{1}{125} + \frac{1}{300} = \frac{17}{1500}$$

$$P(Identical\ twins/twins) = \frac{\frac{1}{300}}{\frac{17}{1500}} = \frac{5}{17}$$

part 2



יש שתי קערות של עוגיות. בקערה 1 יש 10 עוגיות שקדים ו-30 עוגיות שוקולד. בקערה 2 יש
20 עוגיות שקדים ו-20 עוגיות שוקולד. אריק בחר קערה **באקראי** ובחר ממנה עוגיה **באקראי**.
העוגיה שנבחרה היא שוקולד. מה ההסתברות שאריק בחר את קערה 1?

We are looking for P(bowl A/chocolate cookies)=?
Cookie bowl A- 10 almond cookies, 30 chocolate cookies.
Cookie bowl B- 20 almond cookies, 20 chocolate cookies.

$$P(bowl\ A) = \frac{1}{2}$$

$$P(bowl\ B) = \frac{1}{2}$$

$$P(c\ cookies/bowl\ A) = \frac{30}{40} = \frac{3}{4}$$

$$P(c\ cookies/bowl\ B) = \frac{20}{40} = \frac{1}{2}$$

$$P(chocolate\ cookies) = P(c\ cookies/bowl\ A) * P(bowl\ A) + P(c\ cookies/bowl\ B) * P(bowl\ B)$$

$$P(chocolate\ cookies) = \frac{3}{8} * \frac{1}{2} + \frac{1}{4} * \frac{1}{2} = \frac{5}{8}$$

$$P(bowl\ A\ /\ chocolate\ cookies) = \frac{P(c\ cookies/bowl\ A) * P(bowl\ A)}{P(chocolate\ cookies)}$$

$$P(bowl\ A\ /\ chocolate\ cookies) = \frac{\frac{3}{4} * \frac{1}{2}}{\frac{5}{8}} = \frac{3}{5}$$

Question 2

בשנת 1995 חברת M&M הוסיפה את הצבע כחול. לפני השנה הזו, התפלגות הצבעים בשקית M&M נראית כך:

30% Brown, 20% Yellow, 20% Red, 10% Green, 10% Orange, 10% Tan

החל משנת 1995, ההתפלגות נראית כך:

24% Blue , 20% Green, 16% Orange, 14% Yellow, 13% Red, 13% Brown.

לחבר שלכם יש 2 שקיות M&M, אחת משנת 1994 ואחת משנת 1996 והוא לא מוכן לגלות לכם איזו שקית שייכת לאיזו שנה. אבל הוא נותן לכם סוכריה אחת מכל שקית. סוכריה אחת היא צהובה ואחת היא ירוקה. מה הסיכוי שהסוכריה הצהובה הגיעה מהשקית של 1994?

	brown	*red*	*orange*	*yellow*	*green*	*tan*	*blue*
M&M's 1994	*30%*	*20%*	*10%*	*20%*	*10%*	*10%*	*0%*
M&M's 1996	*13%*	*13%*	*16%*	*14%*	*20%*	*0%*	*24%*

We are looking for P(1994/yellow) =?

$$p(yellow) = P(yellow/1994) * P(1994) + P(yellow/1996) * P(1996)$$

$$p(yellow) = \frac{20}{100} * \frac{1}{2} + \frac{14}{100} * \frac{1}{2} = \frac{17}{100}$$

$$p(green) = P(green/1994) * P(1994) + P(green/1996) * P(1996)$$

$$p(green) = \frac{10}{100} * \frac{1}{2} + \frac{20}{100} * \frac{1}{2} = \frac{3}{20}$$

$$P(1994/yellow) = \frac{p(yellow/1994) * P(1994)}{P(yellow)} + \frac{p(green/1996) * P(1996)}{P(green)}$$

$$P(1994/yellow) = \frac{\frac{1}{10} * \frac{1}{2}}{\frac{17}{100}} + \frac{\frac{1}{10} * \frac{1}{2}}{\frac{3}{20}} = \frac{20}{51}$$

Question 3

הלכת לדוקטור בעקבות ציפורן חודרנית. הדוקטור בחר בך **באקראי** לבצע בדיקת דם הבודקת
שפעת חזירים. ידוע סטטיסטית ששפעת זו פוגעת ב-1 מתוך 10,000 אנשים באוכלוסייה. הבדיקה
מדויקת ב-99 אחוז במובן שההסתברות ל false positive היא 1%. הווה אומר שהבדיקה סיווגה
בטעות אדם בריא כאדם חולה היא 1 אחוז. ההסתברות ל- false negative היא 0 – אין סיכוי
שהבדיקה תגיד על אדם החולה בשפעת חזירים שהוא בריא. בבדיקה יצאת חיובי (יש לך שפעת).

א. מה ההסתברות שיש לך שפעת חזירים?

ב. נניח שחזרת מתאילנד לאחרונה ואתה יודע ש-1 מתוך 200 אנשים שחזרו לאחרונה מתאילנד,
חזרו עם שפעת חזירים. בהינתן אותה סיטואציה כמו בשאלה א, מה ההסתברות (המתוקנת)
שיש לך שפעת חזירים?

Part 1

We are looking for P(sick/positive) =?

$$P(sick) = \frac{1}{10000}$$

$$P(positive/sick) = 1$$

$$P(positive/healthy) = \frac{1}{100}$$

$$P(positive) = P(positive/sick) * P(sick) + P(positive/healthy) * P(healthy)$$

$$P(positive) = 1 * \frac{1}{10000} + \frac{1}{100} * \left(1 - \frac{1}{10000}\right) = 0.010099$$

$$P(sick/positive) = \frac{P(positive/sick) * P(sick)}{P(positive)}$$

$$P(sick/positive) = \frac{1 * \frac{1}{10000}}{0.010099} = \frac{100}{10099} = 0.0099$$

Part 2

We are looking for P(sick/positive) =?

$$P(sick) = \frac{1}{200}$$

$$P(positive/sick) = 1$$

$$P(positive/healthy) = \frac{1}{100}$$

$$P(positive) = P(positive/sick) * P(sick) + P(positive/healthy) * P(healthy)$$

$$P(positive) = 1 * \frac{1}{200} + \frac{1}{100} * \left(1 - \frac{1}{200}\right) = \frac{299}{20000}$$

$$P(sick/positive) = \frac{P(positive/sick) * P(sick)}{P(positive)}$$

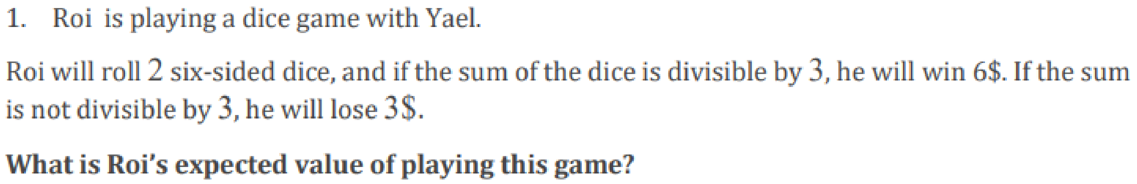
$$P(sick/positive) = \frac{1 * \frac{1}{200}}{\frac{299}{20000}} = \frac{100}{299} = 0.3344\ldots$$

Introduction to data science final project

Part 1

Random variables

Question 1



1. Roi is playing a dice game with Yael.

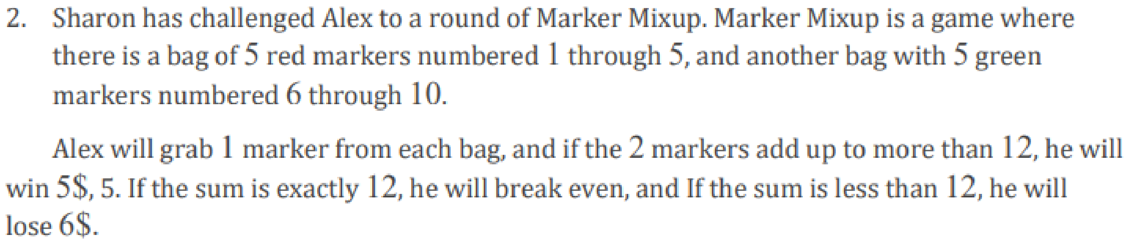
Roi will roll 2 six-sided dice, and if the sum of the dice is divisible by 3, he will win 6$. If the sum is not divisible by 3, he will lose 3$.

What is Roi's expected value of playing this game?

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$$\frac{1}{3} of\ the\ sums\ are\ divisable\ by\ 3. \left(\frac{12}{36} = \frac{1}{3}\right)$$
$$\frac{2}{3} of\ the\ sums\ are\ not\ divisable\ by\ 3. \left(\frac{24}{36} = \frac{2}{3}\right)$$
$$thus\ we\ can\ see\ that\ Roi's\ value\ of\ playing\ the\ game\ is\ 0.$$
$$because\ \frac{2}{3}\ of\ the\ time\ Roi\ will\ lose\ 3\$\ and\ \frac{1}{3}\ of\ the\ time\ Roi\ will\ win\ 6\$.$$
$$\frac{1}{3} * 6\$ + \frac{2}{3} * -3\$ = 0\$$$

Question 2



2. Sharon has challenged Alex to a round of Marker Mixup. Marker Mixup is a game where there is a bag of 5 red markers numbered 1 through 5, and another bag with 5 green markers numbered 6 through 10.

Alex will grab 1 marker from each bag, and if the 2 markers add up to more than 12, he will win 5$, 5. If the sum is exactly 12, he will break even, and If the sum is less than 12, he will lose 6$.

+	6	7	8	9	10
1	7	8	9	10	11
2	8	9	10	11	12
3	9	10	11	12	13
4	10	11	12	13	14
5	11	12	13	14	15

$\frac{6}{25}$ *of the sums are bigger than* 12.

$\frac{15}{25}$ *of the sums are smaller than* 12.

$\frac{4}{25}$ *of the sums are smaller than* 12.

thuswecanseethatAlex'svalueofplayingthegameis − 60.

because $\frac{15}{25}$ *of the time Allex will lose* 6$, $\frac{6}{25}$ *of the time Alex will win* 5$

and $\frac{4}{25}$ *Alex dosent lose or win*. .

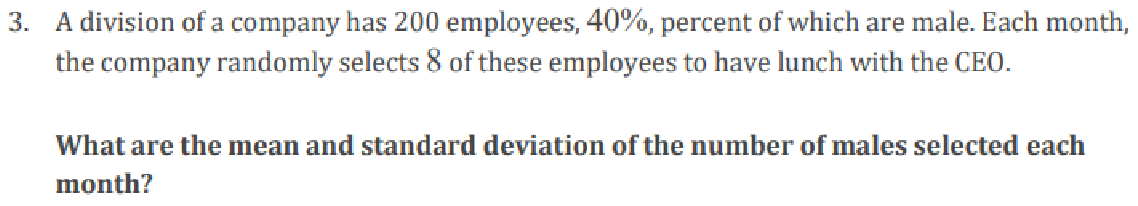
out of every 25 *games Alex plays, he should lose* 15 *times thus he will lose* 15 ∗ 6$ = 90$.

out of every 25 *games Alex plays, he should win* 6 *times thus he will lwin* 6 ∗ 5$ = 30$.

out of every 25 *games Alex plays, he should not win or lose* 4 *times and thus not pay or get payed*.

we can see that the total value of the game is 30$ − 90$ = −60$.

Question 3



3. A division of a company has 200 employees, 40%, percent of which are male. Each month, the company randomly selects 8 of these employees to have lunch with the CEO.

What are the mean and standard deviation of the number of males selected each month?

$P(female) = \frac{300}{500} = \frac{3}{5}$

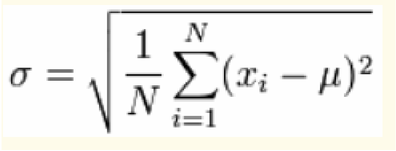
$P(male) = \frac{200}{500} = \frac{2}{5}$

$the\ mean\ is\ 8 * \frac{2}{5} = 3.2$

$this\ number\ is\ 40\%\ of\ the\ 8\ pepole\ who\ go\ to\ lunch\ with\ the\ CEO.$

The std is the 2.7

I calculated it using this formula.



$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

I will explain the formula.

Step 1: do the sigma of: the values of the range of numbers that gave you the mean minus the mean all squared.
Step 2: divide by the number of numbers in your range.
Step 3: do the square root of that number.

$$step\ 1\colon \sum_{i=0}^{8}(i - 3.2)^2 \ = \ 65.76$$

$$step\ 2\colon \frac{65.76}{9} = \frac{548}{75}$$

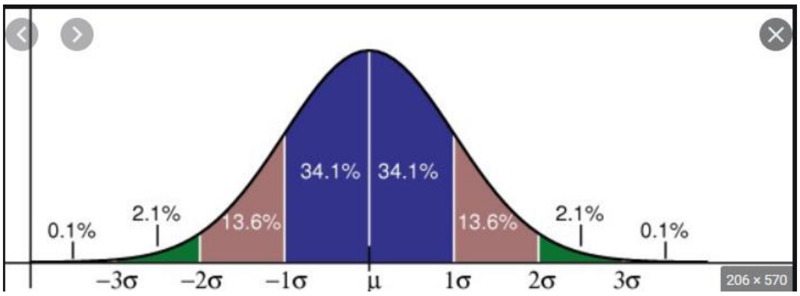
$$step\ 3\colon \sqrt{\frac{548}{75}} = 2.7$$

Question 4

4. Different dealers may sell the same car for different prices. The sale prices for a particular car are normally distributed with a mean and standard deviation of 26,000$ and 2,000$, respectively. Suppose we select one of these cars at random. Let X = the sale price (in thousands of dollars) for the selected car.

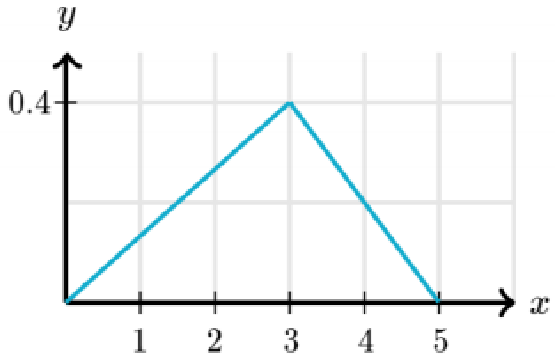
   Find P(26<X<30),

To find the probability of x being in the ranges of 26000<x<30000. We will look at a normally distributed bell curve. In our case the std is 2000, so we want to know what is the probability that x is within 2 std's to the right of the mean. By looking at the graph below we can see that in a normal distribution the area under the graph for 2 std's is 34.1+13.6=47.7%.
thus there is a 47.7% chance that x is in the range 26<x<30.



Question 5

5. Given the following distribution, what is P(x>3)?



To find the probability of x we will calculate the area under the curve.
Area 1= from 0 to 3.

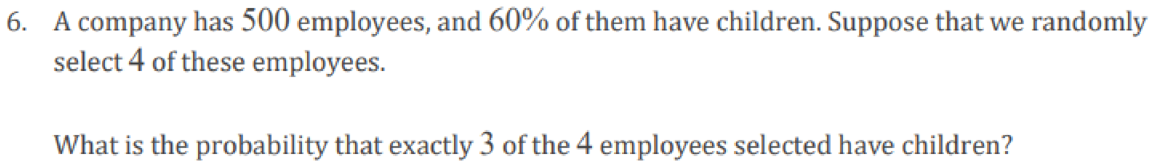
$$area\ 1\ = \frac{3 \ast 0.4}{2} = \frac{3}{5}$$

Area 2=from 3 to 5.

$$area\ 2 = \frac{2 \ast 0.4}{2} = \frac{2}{5}$$

We can see that the sum of both areas is the whole area under the curve. Thus, the probability of x being bigger than 3 is 0.4.

Question 6



6. A company has 500 employees, and 60% of them have children. Suppose that we randomly select 4 of these employees.

What is the probability that exactly 3 of the 4 employees selected have children?

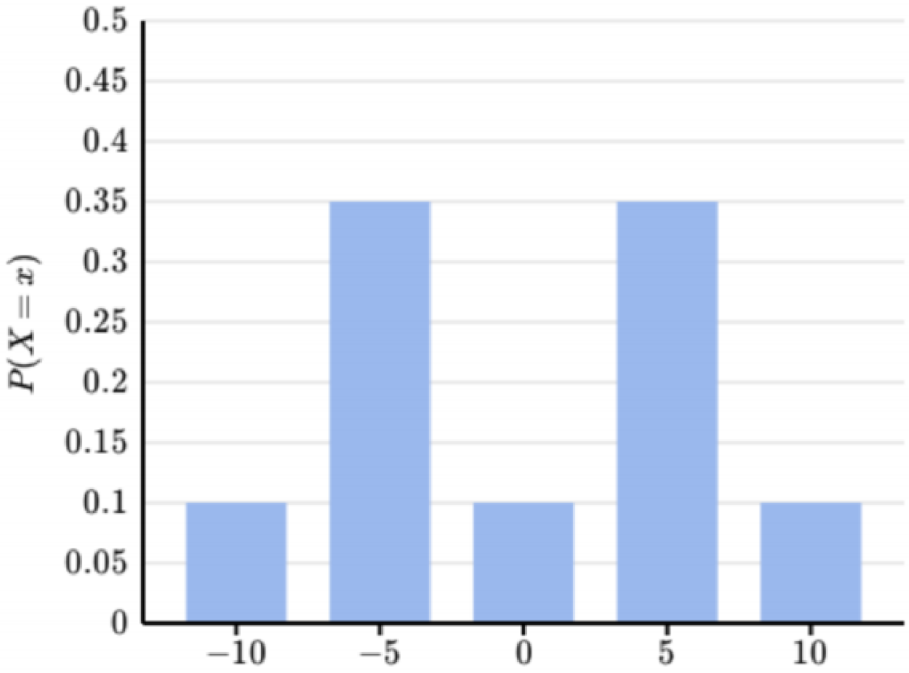
$$P(employees\ with\ children) = \frac{300}{500} = \frac{3}{5}$$

$$P(employees\ without\ children) = \frac{200}{500} = \frac{2}{5}$$

$$\frac{300}{500} * \frac{299}{499} * \frac{298}{498} * \frac{200}{497} * 4 = 0.3463$$

the probability that 3 out of 4 selected employees have children is 0.3463

Question 7



7. Look at the next Graph. What is the expected value of X?

the expected value of x is 0.
using P(X=x). we will calculate x's value.

$$-10 * 0.1 \pm 5 * 0.35 + 0 * 0.1 + 5 * 0.35 + 10 * 0.1 = -1 - 1.75 + 0 + 1.75 + 1 = 0$$