

## סטטיסטיקה למדעי המחשב – תרגיל בית 1

### שאלה 1 (20 נקודות)

ענו נכון/לא נכון עם נימוק קצר.

- א. ניתן לחשב שונות של סכום משתנים מקריים ללא ידיעת השונות המשותפת שלהם. ✓
- ב. ניתן לחשב תוחלת של סכום משתנים מקריים ללא ידיעת השונות המשותפת שלהם. ✓
- ג. אם משתנה מקרי  $Y$  תלוי ב  $X$ , בהכרח מתקיים ש:  $\text{Cov}(X, Y) > 0$ . ✓
- ד. האם ייתכן שההסתברות המותנית של  $A$  בהינתן  $B$  תהיה גדולה יותר מההסתברות השולית למאורע  $A$ ? ✓
- ה. יהיו שלושה מאורעות  $A, B, C$ , ונניח כי  $P(A|B, C) \geq P(B|A, C) > 0$  אז נובע מכך בהכרח ש-  
 $P(A|C) \geq P(B|C)$  ✓

### הערה לשאלות 2,3 (ושאלות עתידיות המערבות שימוש ב-python):

יש להציג את הקוד ואת הפלטים המתקבלים בהרצת הקוד ולא רק את הקוד שכתבתם.

## סתרון שלמה 1

א. דבר נכון. דבר הנכונה דבר שכלל  $\{X_i\}_{i=1}^n$  סדרה של נ"נ  
 משתנים נ"נ: עברו

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$$

אכן (ה'שלב תלוי ב'קיום' השלילי) המשפט  
 (אם ה'שלב יוקדים נכונה  $\{X_i\}_{i=1}^n$  בלתי משתנים, ה'יה נ"נ)  
 $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$  ואם לא ה'יה צורך ב'כיום' (א'ל)

ב. נכון. דבר  $\{X_i\}_{i=1}^n$  נ"נ משתנים נ"נ (א'ל)

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i]$$

וה'חלק א' נ"נ ניתן למד י'עיד משך ה'יה ב'ל' ולכן צורך  
 דפוס על השלילי המשפט

ג. לא נכון. נחמיט בקומה ה'ל:

X - סכום ה'ל' של ש' קוביות

Y - מ'פ'קטור א'ל: "ה'ל' ה'ל' '2 ל' 2"

$$P(X=12, Y=2) = 0 \quad P(Y=1) = \frac{1}{6} \quad P(X=12) = \frac{1}{36}$$

$$\neq \frac{1}{6} \cdot \frac{1}{36} = P(Y=1) \cdot P(X=12) \Rightarrow Y \neq X$$

יכול

$$\text{Cor}(X, Y) = E[XY] - E[X]E[Y] = \#$$

$$E[X] = E[X_1 + X_2] = E[X_1] + E[X_2] = 2 \cdot \frac{7}{2} = 7$$

$\downarrow$   $X = X_1 + X_2$   $\xrightarrow{\text{התכנסות}}$   $\downarrow$   $E[X_1] + E[X_2]$   
 תוצאה הסכום  $\rightarrow$  תוצאת הסכום  $\rightarrow$  תוצאת הסכום

$$E[Y] = P(Y=1) = \frac{1}{6}$$

$\downarrow$   $\frac{1}{3}$   $\frac{1}{6}$

$$E[XY] = E[XY | Y=0] P(Y=0) + E[XY | Y=1] P(Y=1)$$

$\downarrow$   $\frac{1}{3}$   $\frac{1}{6}$

$$= 0 + E[X | Y=1] \cdot \frac{1}{6} = \frac{1}{6} E[2 + X_2] =$$

$\xrightarrow{\text{התכנסות}}$

$$= \frac{1}{6} (2 + E[X_2]) = \frac{1}{6} (2 + \frac{7}{2}) = \frac{2}{6} + \frac{7}{12} = \frac{4}{12} + \frac{7}{12} = \frac{11}{12}$$

$$\text{Cor}(X, Y) = \# = \frac{11}{12} - \frac{1}{6} \cdot 7$$

יכול

$$= \frac{11}{12} - \frac{7}{6} = \frac{11}{12} - \frac{14}{12} = -\frac{3}{12} = -\frac{1}{4} < 0$$

"2 ו-3": תוצאות  $A$  ו- $B$   $\rightarrow$   $P(A) = \frac{1}{6}$   $\rightarrow$   $P(B) = \frac{1}{2}$

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3} > \frac{1}{6} = P(A)$$

ה. יהיו שלושה מאורעות  $A, B, C$  ונניח כי  $P(A|B, C) \geq P(B|A, C) > 0$  אז נובע מכך בהכרח ש-

$$P(A|C) \geq P(B|C)$$

מחמת זה נובע:

$$\frac{P(A, B, C)}{P(B, C)} \geq \frac{P(B, A, C)}{P(A, C)} > 0$$

$$\Rightarrow \frac{P(A, C)}{P(B, C)} \geq 1 \quad (*)$$

$$P(A, B, C) = P(B, A, C) > 0$$

$$\frac{P(A, C)}{P(C)} \geq \frac{P(B, C)}{P(C)} \quad \text{כלומר} \quad P(A|C) \geq P(B|C) \quad \text{כלומר}$$

$$(*) \quad P(A, C) \geq P(B, C) \quad \text{כלומר}$$



שאלה 2 (30 נקודות)

בשאלה זו, מומלץ להשתמש ב-pyplot ו-numpy. תוכלו ליבא אותן באמצעות הפקודות הבאות:

```
import matplotlib.pyplot as plt
```

```
import numpy as np
```

בנוסף, יבאו מתוך scipy.stats את binom:

```
from scipy.stats import binom
```

להלן פירוט פונקציות שעשויות להיות שימושיות:

מחזירה תצפיות מתוך התפלגות בינומית. לדוגמא, הרצת:  binom.rvs(n=20, p=0.2, size=10)  תחזיר וקטור של 10 תצפיות מהתפלגות $\text{Bin}(n = 20, p = 0.2)$ .	binom.rvs(n,p,size)
מחזירה ערכים של פונקציית ההתפלגות המצטברת של  $\text{Bin}(n = n, p = p)$ עבור הערכים שהוזנו בוקטור k.	binom.cdf(k, n, p)

פונקציות שימושיות נוספות:

- הפונקציה plt.scatter(x,y) מציירת גרף פיזור עבור וקטורים x ו-y-הזהים באורכם.
- הפונקציה plt.plot(w,z) מציירת קו רציף בין הנק' של x ו-y (יכולים להיות שונים מאלו ב scatter).
- כדי להציג מספר תרשימים בבת אחת, ניתן להריץ מספר שורות ולאחר מכן plt.show():

```
plt.scatter(x,y)
```

```
plt.plot(w,z)
```

```
plt.show()
```

א. בנו פונקציה בשם `Empirical_F()` המקבלת וקטור  $X$  ומחשבת את הערך של פונקציית ההתפלגות המצטברת האמפירית (לפי  $X$ ) לכל אחד מהערכים בוקטור. הפונקציה מחזירה מטריצה עם עמודה  $x$  הכוללת את הווקטור המקורי, ועמודה  $f$  בה ערך פונקציית ההתפלגות המצטברת האמפירית. לדוגמא:

```
x=np.array([8,2,-13,4,-9,0,18,4,-5,10,1,-7,7,13,-5,-16,-9,18,-10,0])
a=Empirical_F(x)
print(a)
```

[-16.	0.05]
[-13.	0.1 ]
[-10.	0.15]
[-9.	0.25]
[-9.	0.25]
[-7.	0.3 ]
[-5.	0.4 ]
[-5.	0.4 ]
[ 0.	0.5 ]
[ 0.	0.5 ]
[ 1.	0.55]
[ 2.	0.6 ]
[ 4.	0.7 ]
[ 4.	0.7 ]
[ 7.	0.75]
[ 8.	0.8 ]
[10.	0.85]
[13.	0.9 ]
[18.	1. ]
[18.	1. ]]

- ב. צרו וקטור  $X$  של 20 תצפיות מההתפלגות `Bin(5,1/6)`.
- ג. חשבו את פונקציית ההתפלגות המצטברת האמפירית של האיברים ב-  $X$  בעזרת הפונקציה שיצרתם בסעיף א.
- ד. ציירו גרף של פונקציית ההתפלגות המצטברת שחושבה בסעיף ג'. הגדירו את הגבול של ציר ה- $y$  להיות בין 0 ל-1 (היעזרו ב- `plt.ylim`, וכן ב- `plt.step`)
- ה. צרו וקטור  $Y$  המקבל את כל הערכים הבדידים מ-0 עד 5 כולל, וחשבו את ערכי פונקציית ההתפלגות המצטברת של התפלגות `Bin(5,1/6)` (ההתפלגות התאורטית).
- ו. הוסיפו לגרף מסעיף ד' את הקו המתאר את ההתפלגות המצטברת שחושבה בסעיף ה'.
- ז. השוו את ההתפלגות האמפירית שנגזרת מהווקטור  $X$  עם ההתפלגות התיאורטית. הגדילו את מספר התצפיות של סעיף ב' מ-20 ל-100,200,1000. מה המסקנה לגבי הקשר בין גודל המדגם, ההתפלגות האמפירית וההתפלגות התיאורטית?
- בנוס (5 נקודות):** חזרו על סעיפים ב' עד ה' עם התפלגות פואסונית עם פרמטר 2. מצאו את הפונקציות הפואסוניות המקבילות לפונקציות הבינומיות.

Assignment\_1.ipynb

File Edit View Insert Runtime Tools Help [Cannot save changes](#)

+ Code + Text Copy to Drive

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import binom
import sys
import numpy
numpy.set_printoptions(threshold=sys.maxsize)
```

```
[36] x=np.array([8,2,-13,4,-9,0,18,4,-5,10,1,-7,7,13,-5,-16,-9,18,-10,0])
def Empirical_F(x):
    """Calculates the empirical CDF for a given probability space.

    Parameters
    -----
    x : np.array
        The given probability space (default is None)

    Returns
    -----
    np.array
        A matrix with two columns. The first is the value from x and
        the second column holds the corresponding CDF
        for that value.

    """
    n=len(x)
    d=dict()
    cdf=0
    for w in sorted(x):
        if d.get(w) is None:
            d[w]=1
        else:
            d[w]=d[w]+1
    w = list()
    f = list()
    for value,occurrences in d.items():
        cdf += occurrences / n
        for i in range(occurrences):
            w.append(value)
            f.append(cdf)
    res = np.array([w,f]).transpose()
    return res
```

. / c

```
a=Empirical_F(x)
print(a)
```

```
[[-16.    0.05]
 [-13.    0.1 ]
 [-10.    0.15]
 [ -9.    0.25]
 [ -9.    0.25]
 [ -7.    0.3 ]
 [ -5.    0.4 ]
 [ -5.    0.4 ]
 [  0.    0.5 ]
 [  0.    0.5 ]
 [  1.    0.55]
 [  2.    0.6 ]
 [  4.    0.7 ]
 [  4.    0.7 ]
 [  7.    0.75]
 [  8.    0.8 ]
 [ 10.    0.85]
 [ 13.    0.9 ]
 [ 18.    1.  ]
 [ 18.    1.  ]]
```

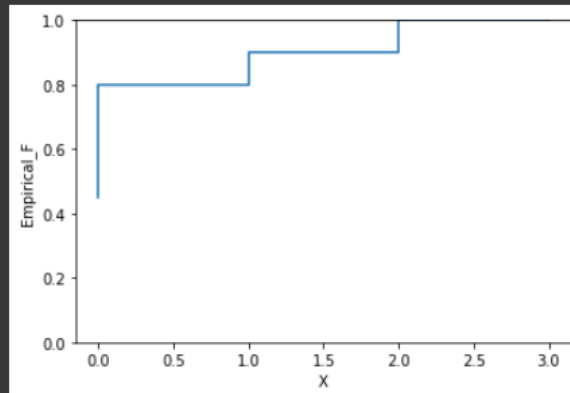
```
X=binom.rvs(n=5,p=1/6,size=20)
```

```
[46] Empirical_F(X)
```

```
array([[0. , 0.4 ],
       [0. , 0.4 ],
       [0. , 0.4 ],
       [0. , 0.4 ],
       [0. , 0.4 ],
       [0. , 0.4 ],
       [0. , 0.4 ],
       [1. , 0.75],
       [1. , 0.75],
       [1. , 0.75],
       [1. , 0.75],
       [1. , 0.75],
       [1. , 0.75],
       [1. , 0.75],
       [2. , 0.95],
       [2. , 0.95],
       [2. , 0.95],
       [2. , 0.95],
       [3. , 1.  ]])
```

c + n

```
[87] plt.xlabel('X')
      plt.ylabel('Empirical_F')
      plt.ylim(0, 1)
      plt.step(x=bin[:,0],y=bin[:,1])
      plt.show()
```

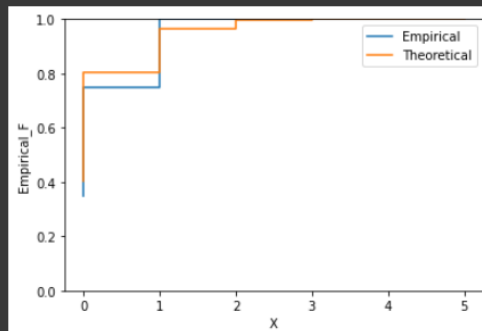


```
[22] Y=np.array([0,1,2,3,4,5])
      binom.cdf(Y,5,1/6)

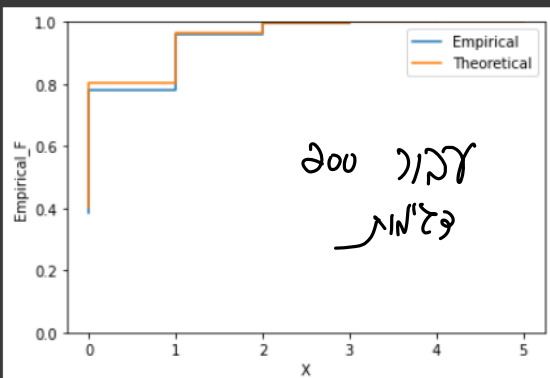
array([0.40187757, 0.80375514, 0.96450617, 0.99665638, 0.9998714 ,
        1.          ])
```

```
[24] Y=np.array([0,1,2,3,4,5])
      theoretical = binom.cdf(Y,5,1/6)
```

```
[28] plt.xlabel('X')
      plt.ylabel('Empirical_F')
      plt.ylim(0, 1)
      plt.step(x=bin[:,0],y=bin[:,1])
      plt.step(x=Y, y=theoretical)
      plt.legend(['Empirical', 'Theoretical'])
      plt.show()
```

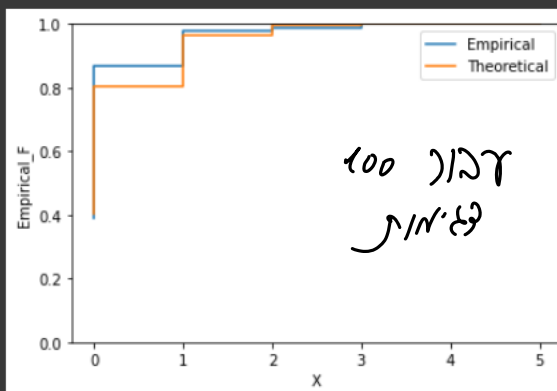






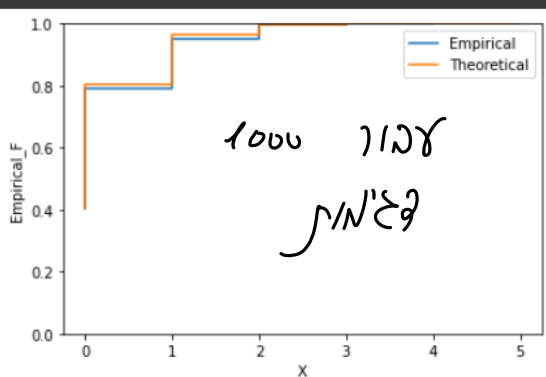
דעם  
פאנא

5



100 127  
5/11/89

```
plt.xlabel('X')
plt.ylabel('Empirical_F')
plt.ylim(0, 1)
plt.step(x=bin[:,0],y=bin[:,1])
plt.step(x=Y, y=theoretical)
plt.legend(['Empirical', 'Theoretical'])
plt.show()
```



1000 7128  
J/N'29

2) 100%

כפי ש'אמך כראוי, נכל שלט הגרעסור חבדו הירחפלאיוור  
הי אמה'כז — והיה אלהים סומכוס דהגאנפיות.

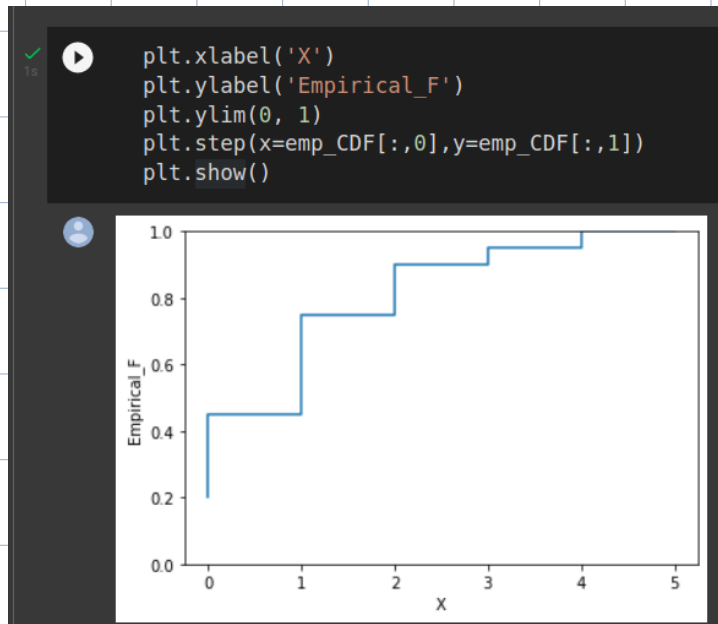
0110

2

```
[23] mu=2
      X=poisson.rvs(mu,size=20)
      print(X)

[1 1 3 3 2 3 1 2 2 4 0 1 3 1 2 1 2 2 4 1]
```

ع + 3



```
print(emp_CDF)

[[0.  0.2 ]
 [0.  0.2 ]
 [0.  0.2 ]
 [0.  0.2 ]
 [1.  0.45]
 [1.  0.45]
 [1.  0.45]
 [1.  0.45]
 [1.  0.45]
 [1.  0.45]
 [2.  0.75]
 [2.  0.75]
 [2.  0.75]
 [2.  0.75]
 [2.  0.75]
 [2.  0.75]
 [2.  0.75]
 [2.  0.75]
 [3.  0.9 ]
 [3.  0.9 ]
 [3.  0.9 ]
 [3.  0.9 ]
 [4.  0.95]
 [5.  1.  ]]
```

```
import array
Y=np.array([0,1,2,3,4,5])
theoretical = poisson.cdf(k=Y,mu=mu)
print(theoretical)

[0.13533528 0.40600585 0.67667642 0.85712346 0.94734698 0.98343639]
```

2

### שאלה 3 (25 נקודות)

בשאלה זו השתמשו ב-pandas:

```
import pandas as pd
```

הורידו את הקובץ `appendicitis.csv` מהמודל ושמרו אותו בתיקייה כלשהי (לצורך הדוגמא, התיקייה

היא: `c:/the path we chose/appendicitis.csv`). העלו את הקובץ הזה לפייתון באופן הבא:

```
path = "c:/the path we chose/appendicitis.csv"
```

```
df = pd.read_csv(path)
```

רצוי להשתמש בפונקציה `df.head()` על מנת להכיר את הקובץ איתו אתם עובדים.

בטבלה נתונים של חולים שנשלחו לניתוח להסרת תוספתן בבית החולים מאיר במהלך תקופה של שנתיים. לאחר הניתוח שולחים את הרקמה לבדיקת פתולוגיה על מנת לבדוק האם באמת היה צורך בניתוח או לא. בעמודה בשם "Pathology" הערך "1" מייצג ניתוח מוצדק ו-"2" מייצג חולה בריא שנשלח לניתוח לשווא. העמודות `Age` ו-`Sex` מייצגות את המין והגיל של המטופלים בהתאמה.

- מהי ההסתברות האמפירית של מנותח להתגלות כמנותח שווא? ✓
- הציגו טבלת שכיחות הסופרת את מספר הפריטים בשילוב פתולוגיה ומין המטופל. (ניתן להיעזר ב-`df.groupby`) ✓
- מה ההסתברות האמפירית להישלח לניתוח שווא עבור גבר? מה ההסתברות האמפירית להישלח לניתוח שווא עבור אישה? האם ניתן לקבוע שמגדר מסוים יותר סביר להישלח לניתוח שווא באוכלוסייה הכללית? ✓

✓ **בנוס (5 נקודות):** הציגו גרף של פונקציית ההתפלגות המצטברת האמפירית של גיל הגברים שנשלחו לניתוח הסרת תוספתן ושל גיל הנשים. שני הקווים צריכים להופיע באותו גרף ולהיות בעלי צבעים שונים. השתמשו בפונקציות `title`, `ylabel`, `xlabel` של `pyplot` על מנת לתת כותרות לגרפים. השתמשו בפונקציית `legend` של `pyplot` על מנת להוסיף מקרא.

פתרון :  
 נוסח לפי ההסתברות :

$P(2)$   
 //

0.24344569288389514

```
[76] import pandas as pd
df=pd.read_csv("/content/appendicitis.csv")
df=pd.DataFrame(df)
df.head()
```

	Sex	Age	Pathology
0	Male	72	1
1	Female	88	1
2	Male	74	1
3	Female	69	1
4	Female	69	2

```
def Empirical_P(X, w):
    """Calculates the empirical probability for a given probability space.

    Parameters
    -----
    x : np.array
        The given probability space (default is None)
    w : float
        The value to calculate empirical probability for

    Returns
    -----
    float
        The empirical P(w)

    """
    size = 0
    w_count = 0
    for val in X:
        size+=1
        if val==w:
            w_count+=1
    return w_count/size
Empirical_P(df["Pathology"],2)
```

0.24344569288389514

```
[85] df.groupby(["Sex", "Pathology"]).count()
```

	Sex	Pathology	
Female	1	101	
	2	47	
Male	1	101	
	2	18	

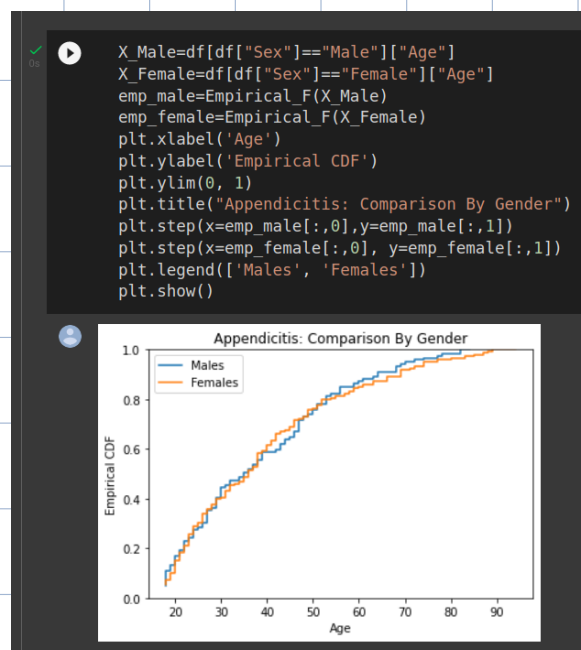
$$P(2 | male) = \frac{P(2, male)}{P(male)} = \frac{\frac{18}{267}}{\frac{119}{267}} = \frac{18}{119} = 0.1512605$$

$$P(2 | female) = \frac{P(2, female)}{P(female)} = \frac{\frac{47}{267}}{\frac{148}{267}} = \frac{47}{148} = 0.3175675$$

למחזור, הסיכוי של אישה תישאר עניינת שוא  
 גבוה יותר מהסיכוי של גבר להישאר עניינת צומח

עם זאת, כפי שהאנוכי הכירה השיקולים הברורים של  
 סימסון לא ניתן לקבוע תג-משמעותי שיש נושא להישאר  
 יותר אנטישמי שוא מאשר גברים. יתכן שהוכחה של פולמר נוסף  
 אנטישמי, עלמל מפעם ביאוגי, תשנו את המסקנה. יתכן כי הנוסחה  
 להישאר עניינת שוא קווקא נכונה למצב הביאוגי גבוה ויותר  
 נשים שסקרי נמצאו במצב זה ואכן זה אף שבתוך אל מצב  
 הביאוגי, אחוז וקשים שגאות אנטישמי שוא נמצא מחלה של  
 נוגדים, ואכן אי-דקיות של גזות ובה בה שנין עלולה להיות למסקנה  
 שליה.

היוונים



שאלה 4 (25 נקודות) ✓

ניזכר בטבלה השכיחות של הישרדות נוסעי הטיטניק (משתנה Y) בשילוב עם סוג המחלקה (X):

		Class				Total
		First	Second	Third	Crew	
Survival	Alive	202	118	178	212	710
	Dead	123	167	528	673	1491
	Total	325	285	706	885	2201

נניח שמתווסף מידע לגבי גיל הנוסעים (משתנה Z) כאשר הערכים האפשריים הינה Young או Old. המידע מסוכם בטבלה הבאה\*:

		Class and Age								
		First		Second		Third		Crew		Total
		Y	O	Y	O	Y	O	Y	O	
Survival	Alive	10	192	100	18	170	8	210	2	710
	Dead	3	120	140	27	500	28	665	8	1491
	Total	13	312	240	45	670	36	875	10	2201

\*נתוני הגיל מפוברקים לטובת התרגיל.

מצאו והסבירו כיצד בא לידי הפרדוקס של סימפסון מתוך הנתונים בטבלה המכילה את הנתונים לגבי גיל הנוסעים. ספקו הוכחה חישובית לפרדוקס וספקו הסבר מילולי שמצדיק את קיומו של הפרדוקס במקרה זה.

## פתרון 4

ננתן את סיכויי ההיכשלות של יורש

בהתאם לגילו.  $Y=1$  (יורש,  $Y=0$  (לא יורש).

נרצה לדעת  $P(Y=0 | Z=old)$

$$P(Y=0 | Z=old) = \frac{P(Y=0, Z=old)}{P(Z=old)} = \frac{\frac{183}{2201}}{\frac{403}{2201}}$$

$$= \frac{183}{403} = 0.455$$

$$P(Y=0 | Z=young) = \frac{P(Y=0, Z=young)}{P(Z=young)} = \frac{\frac{1308}{2201}}{\frac{1798}{2201}}$$

$$= \frac{1308}{1798} = 0.727$$

כלומר, לכאורה, המסקנה היא שהסיכוי של בן זוג להינשא  
אם הוא צעיר, הוא גבוה יותר מאשר אם הוא מבוגר.

כעת ננסה לנתח את סוג המסקנה.

עבור המסקנה הזו:

$$\frac{15}{13} \approx 0.77 \text{ מהצפייה של } 0.62 \text{ (הנחה של } 0.62 \text{)}$$

זכור המחלקה השנייה:

$$0.4 \approx \frac{18}{45} \quad \text{מחצית שנייה} \quad 0.42 \approx \frac{5}{12} = \frac{100}{240} \quad \text{מהמבוזרים.}$$

מחלקה שלישית:

$$0.25 \approx \frac{170}{670} \quad \text{מחצית שנייה} \quad 0.22 \approx \frac{8}{36} \quad \text{מהמבוזרים.}$$

$$0.24 = \frac{5}{875} \quad \text{מחצית שנייה} \quad \text{לחצות}$$

$$0.2 = \frac{1}{5} = \frac{2}{10} \quad \text{מחצית} \quad \text{המבוזרים.}$$

כלומר נאמי מכלול על כל מחלקה בתכנה מקבלים מחצית שנייה  
בויקא שרבי יגד מהמבוזרים.

נרצח ערוא שכל מחלקה טובה יגד כך טכני (הה'שנתי)  
של אדם מילי. בתנ"ך, נאמי מכלול (המבוזרים) נאמי  
לשנה במחלקה נאמי יגד ולכן בתנ"ך של  
מכלול המחלקה מהכלל שביכולת שחזר נאמי מכלול של המבוזרים  
אך מכלול הנאמי של צו"רם נאמי מכלול של צו"רם  
לכלל מכלול נאמי יגד.

סוף פתרון ת"ה 1