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, בהכרת מתקיים ש: $\mathrm{Cov}(X,Y) > 0$.

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 <u>ואת הפלטים המתקבלים</u> בהרצת הקוד ולא רק את הקוד שכתבתם.

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:

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

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?

בונוס (5 נקודות): חיזרו על סעיפים ב' עד ה' עם התפלגות פואסונית עם פרמטר 2. מצאו את הפונקציות הפואסוניות המקבילות לפונקציות הבינומיות.

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 Sex ו Age -מייצגות את המין והגיל של המטופלים בהתאמה.

א. מהי ההסתברות האמפירית של מנותח להתגלות כמנותח שווא?

ב. הציגו טבלת שכיחות הסופרת את מספר הפריטים בשילוב פתלוגיה ומין המטופל. (ניתן להיעזר ב-df.groupby)

ג. 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?

בונוס (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 xlabel ,ylabel, title של pyplot על מנת לתת כותרות לגרפים. השתמשו בפונקציית legend של pyplot על מנת להוסיף מקרא.

שאלה 4 (25 נקודות)

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

Survival	Class: First	Second	Third	Crew	Total
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.