מבוא לניתוח נתונים - 096202

מבחן מועד א' – סמסטר חורף תש"ף

טור א'

תאריך הבחינה: 20.02.2020

מרצה: אורי פלונסקי

מתרגל: רפאל שללה

הוראות:

1. לרשותכם **שעתיים וחצי** לפתור את הבחינה.
2. מותר להשתמש במחשבון פשוט בלבד. אין להשתמש בכל חומר עזר אחר.
3. הבחינה כוללת 16 עמודים (כולל עמוד זה) ובהם 12 שאלות. בדקו בתחילת הבחינה שיש ברשותכם את כל העמודים.
4. ניקוד כל שאלה (ולרוב כל סעיף) מצוין לידה. סך כל **הניקוד האפשרי בבחינה הוא 102** נקודות.
5. **את התשובות יש לכתוב רק על גבי טופס הבחינה, ובמקומות המיועדים לכך בלבד.** מחברות הטיוטה יושמדו לאחר סיום הבחינה וממילא לא יבדקו.
6. חובה לכתוב מספר תעודת זהות על כל אחד מדפי טופס הבחינה.
7. אסור להפריד את דפי טופס הבחינה.
8. בשום שלב, במהלך הבחינה או לאחר סיומה, אסור להוציא מחדר הבחינה שום דבר שלא הבאתם איתכם לחדר הבחינה. בפרט, חובה להחזיר בסיום הבחינה את הטופס וכל מחברת בה השתמשתם.

בהצלחה!

ת"ז: _____

<u>שאלה 1 (15 נק')</u>

ענו נכון או לא נכון. נמקו תשובתכם במשפט אחד.

א. כאשר השערת האפס נכונה, לעולם לא נדחה אותה אם נשתמש במבחנים סטטיסטיים בצורה מתאימה. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ב. נתון משתנה המייצג הכנסה ומקבל את הערכים 1, 2, או 3 בלבד, כאשר: 1 = הכנסה מתחת ל-8000 ₪ בחודש; 2 = הכנסה בין 8001 ל-15000 ₪ בחודש; 3 = הכנסה של יותר מ-15000 ₪ בחודש. משתנה זה הינו משתנה קטגוריאלי. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ג. בתהליך K-fold cross validation על אלגוריתם kNN, מחלקים את הנתונים ל-K חלקים ובכל חלק בודקים את הדיוק של שימוש במספר שכנים, k, אחר. לבסוף בוחרים את מספר השכנים k שנותן את הדיוק הגבוה ביותר. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ד. נתון שבאמצעות שיטת בוטסטראפ חושב רווח סמך ברמת ביטחון של 99% להפרש בין מספר הבאגים היומי הממוצע של מתכנת ממחלקה א' למספר הבאגים היומי הממוצע של מתכנת ממחלקה ב' והתקבל [1.3, 2.7]. לכן, ברמת מובהקות של 5%, ניתן לדחות את השערת האפס שמספר הבאגים היומי הממוצע של מתכנת ממחלקה א' זהה למספר הבאגים היומי הממוצע של מתכנת ממחלקה ב'. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

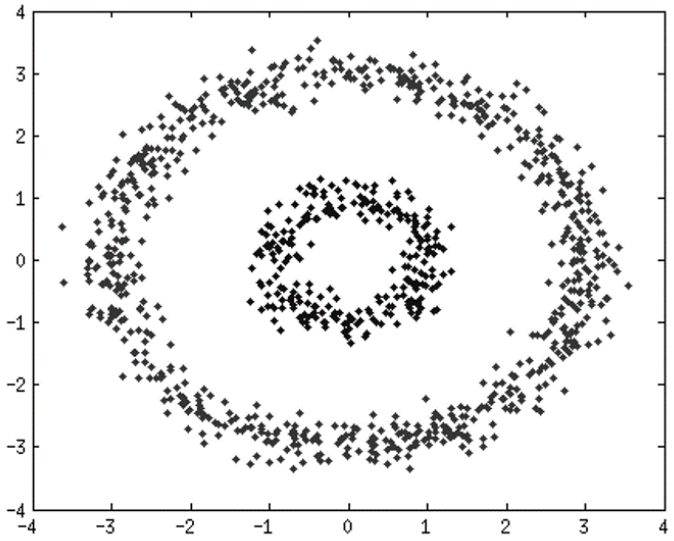
[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ה. בידינו תצפיות עם שני משתנים נומריים. תרשים פיזור של הקשר בין שני המשתנים מוצג להלן. על נתונים אלו הורץ אלגוריתם K-Means עם K=2. האלגוריתם סיווג את כל הנקודות שיוצרות מעגל קטן פנימי לאשכול (cluster) אחד ואת כל הנקודות שיוצרות מעגל גדול חיצוני לאשכול (cluster) שני. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____



<u>שאלה 2 (6 נק')</u>

לפני מבחן במתמטיקה לכיתות ח' בבית ספר "ברושים" נאספו נתונים מכלל תלמידי כיתות ח' בבית הספר בנוגע לסך כל זמן הלמידה שלהם למבחן (בדקות). לאחר מכן, נתונים אלו הושוו לציוני המבחן. התקבל כי הקורלציה (מתאם) הלינארית בין זמן הלמידה (בדקות) לבין ציון המבחן היא 0.74. לכל אחד מההיגדים הבאים כתבו נכון או לא נכון ונמקו במשפט אחד.

א. מאחר שהנתונים כוללים את כל האוכלוסייה (ולא רק מדגם מתוכה), ניתן להסיק שעבור אוכלוסייה זו, זמן למידה ארוך יותר למבחן במתמטיקה גורם לציונים גבוהים יותר במבחן.

_____

_____

ב. באמצעות הנתונים שמתואר שנאספו, ניתן להשתמש בשיטת בוטסטראפ על מנת לחשב רווח סמך ברמת ביטחון 95% לקורולציה הלינארית בין זמן למידה וציון המבחן ואם יתקבל שרווח הסמך לא כולל את הערך 0, נוכל לדחות את ההשערה שאין קורולציה לינארית בין זמן למידה וציון המבחן.

_____

_____

ת"ז: _____

<u>שאלה 3 (3 נק')</u>

עבור איזה מבין השאלות הבאות הכי הגיוני להשתמש באלגוריתם kNN? (הקיפו את התשובה הנכונה)

א. האם לקוחות רשומים מוציאים בממוצע יותר כסף באתר מאשר לקוחות שאינם רשומים?

ב. בהתבסס על היסטורית החיפוש והגלישה של לקוח באתר, האם הוא צפוי לרכוש מוצר כלשהו בשבוע הקרוב?

ג. מהו זמן הגלישה הממוצע של לקוחות רשומים באתר?

ד. איזה מהלקוחות הרשומים באתר מתנהגים בצורה דומה זה לזה מבחינת היסטורית החיפוש והקנייה שלהם מהאתר?

<u>שאלה 4 (7 נק')</u>

בכד אטום מבחוץ יש 3 כדורים אדומים, 2 כדורים ירוקים וכדור אחד שחור. לבד מהצבע, כל הכדורים זהים לחלוטין.

א. במשחק א', שחקן שולף מהכד שני כדורים בזה אחר זה **וללא החזרה**. אם <u>אף אחד</u> משניהם אינו אדום, הוא מנצח במשחק. מה ההסתברות של השחקן לנצח במשחק א'? (2 נק')

_____

_____

_____

ב. במשחק ב', החוקים זהים לאלו של משחק א', אבל כעת, לאחר שליפת כל אחד מהכדורים, הם **מוחזרים לכד** לפני שליפת הכדור הבא. שוב, על מנת לנצח, יש לשלוף 2 כדורים <u>שאינם</u> אדומים. מה ההסתברות לנצח במשחק ב'? (2 נק')

_____

_____

_____

ג. שחקן משחק ראשית את משחק א'. **אם** השחקן מנצח במשחק א', הכדורים ששלף מוחזרים לכד ואז השחקן זכאי לשחק את משחק ב'. **אם** השחקן משחק ומנצח גם במשחק ב', הוא זוכה בפרס. מה ההסתברות ששחקן **<u>לא יזכה</u>** בפרס? (3 נק')

_____

_____

_____

_____

ת"ז: _____

שאלה 5 (16 נק')

נשיא הטכניון טוען שבוגרי טכניון שכסטודנטים למדו את הקורס מבוא לניתוח נתונים הם בעלי סיכוי גבוה יותר לעבוד כמדעני נתונים שנה לאחר סיום התואר מאשר בוגרי הטכניון שכסטודנטים לא למדו את הקורס מבוא לניתוח נתונים. לשם בדיקת הטענה, הוא מבקש מאלון לאסוף נתונים ולערוך מבחן סטטיסטי.

א. מהי האוכלוסייה שלגביה או האוכלוסיות שלגביהן רוצים לבדוק את ההשערות? (2 נק')

_____

_____

ב. הגדירו את השערת האפס ואת ההשערה האלטרנטיבית (3 נק')

_____

_____

_____

_____

ג. מהו סטטיסטי המבחן שבאמצעותו תבדקו את ההשערות? (2 נק')

_____

_____

ד. אלון מצליח להשיג גישה לרשימת מספרי הטלפון של כלל בוגרי הטכניון (שנה לאחר סיום התואר) שלקחו את הקורס מבוא לניתוח נתונים. הוא בוחר באקראי 300 מספרי טלפון ומתקשר אליהם במטרה לשאול האם הם עובדים כיום כמדעני נתונים. מאחר שחלקם לא זמינים וחלקם לא מעוניינים לענות, הוא בסופו של דבר מקבל נתונים של 237 מהם.
הציעו הטיה אחת אפשרית בתהליך איסוף הנתונים המתואר אשר עלולה להשפיע על התוצאות של הניתוח שאלון צפוי לעשות. הסבירו מה הבעיה שעלולה להיווצר. (3 נק')

_____

_____

_____

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ה. בנוסף, אלון מקבל גישה לנתונים מתוך מדגם מקרי פשוט של מדעני נתונים בארץ. נתונים אלו כוללים בין היתר את שנת סיום התואר ומוסד אקדמי בו למדו. האם נתונים האלה, יחד עם הנתונים המתוארים בסעיף ד', מספיקים על מנת לבדוק את הטענות של נשיא הטכניון? אם כן, הסבירו בקצרה כיצד תבדקו את הטענה. אם לא, הסבירו מדוע הנתונים לא מספיקים ואילו נתונים נוספים דרושים כדי לבדוק אותה (הניחו שבנתונים שנאספו בסעיף ד' אין הטיות). (4 נק')

_____

_____

_____

_____

_____

ו. בסופו של דבר אלון מצליח להשיג מספיק נתונים (אלו מסעיפים קודמים ו/או נתונים אחרים) על מנת לבחון כיאות את הטענה של נשיא הטכניון. הוא עורך את המבחן הסטטיסטי (בצורה נכונה) ומסיק שיש לדחות את השערת האפס. מכאן ניתן להסיק שלמידת הקורס מבוא לניתוח נתונים גורמת לעלייה בסיכוי לעבוד כמדען נתונים. נכון או לא נכון? נמקו במשפט אחד. (2 נק')

_____

_____

ת"ז: _____

שאלה 6 (9 נק')

תפקידה של ועדת שחרורים להכריע האם יש לאשר את קיצור זמן מאסרו של אסיר מסוים (לרוב לאחר שריצה שני שלישים מהעונש המוקצב לו) ולשחררו מבית הסוהר (תחת הגבלות מסוימות). לפי חוק, על הועדה לאשר את קיצור זמן המאסר אלא אם היא מתרשמת שיש בכך סכנה לציבור או תחת תנאים אובייקטיבים (למשל אם האסיר לא לקח אחריות על מעשיו).

מחקר שנעשה לאחרונה על החלטותיה של ועדת השחרורים מצא כי אסירים ממוצא ערבי משוחררים באחוזים נמוכים בהרבה מאשר אסירים ממוצא יהודי. כמו כן, המחקר מצא כי אסירים ממוצא ערבי שקיצור עונשם אושר ע"י הועדה מבצעים משמעותית פחות פשעים לאחר שחרורם מאשר אסירים ממוצא יהודי. לפי תוצאות אלו ואחרות, המחקר הסיק כי ישנה הטיה בהחלטות הועדה כנגד אסירים ממוצא ערבי.

לאחר שקראה את המחקר, שרון, שלמדה לאחרונה למידת מכונה, החליטה לפתח אלגוריתם סיווג מתוחכם שמטרתו לנבא האם אסיר שיקוצר עונשו צפוי לבצע פשעים נוספים לאחר שחרורו.

לטענתה, מאחר שהאלגוריתם הוא ישות מתמטית אשר מתוכננת לייצר חיזוי מקסימלי, הוא לא יושפע מהטיות כאלה ואחרות כנגד מי מהאסירים. לכן, האלגוריתם יוכל להחליף את החלטות ועדת השחרורים ולשפר את הדיוק של החיזוי הדרוש (גם ע"י צמצום ההטיה כנגד אסירים ממוצא ערבי).

לשם אימון האלגוריתם, שרון השתמשה בנתוני המחקר המוזכר לעיל יחד עם פיצ'רים שונים כמו גיל, מין, מקום מגורים, מצב משפחתי, משלח יד, סוג העבירה, העונש שהוטל על האסיר, עבר פלילי נוסף וכדומה. על מנת להימנע ממצב שבו האלגוריתם "לומד" להפלות נגד אסירים ממוצא ערבי, שרון לא השתמשה בפיצ'ר "מוצא" לשם החיזוי.

הציגו שתי בעיות שצפויות להתעורר עם האלגוריתם של שרון. התייחסו רק לסיפור לעיל.

_____

_____

_____

_____

_____

_____

_____

_____

_____

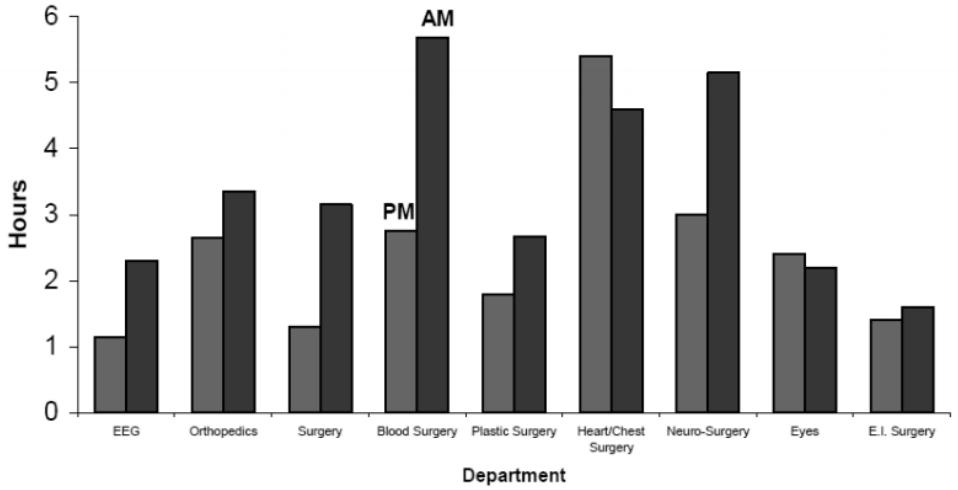
_____

ת"ז: _____

<u>שאלה 7 (7 נק')</u>

על מנת לשפר את היעילות, מנהלת בית חולים "רפואה שלמה" החליטה על תכנית לפיה חדרי הניתוח יהיו פעילים גם בשעות אחר הצהריים. על מנת לייצר תמריצים לרופאים שיישארו לנתח גם בשעות אחה"צ, הוחלט כי בשעות אלו הם יקבלו שכר לפי מספר ניתוחים ולא שכר שעתי כמקובל. כלומר, בעוד שבשעות הבוקר, התשלום לרופאים נעשה לפי שעת עבודה, בשעות אחה"צ, התשלום נעשה לפי מספר ניתוחים שהם מבצעים. לאחר תקופת הרצה של תכנית זו שכללה מאות רבות של ניתוחים, נמדד משך ניתוח ממוצע, בכל אחת מ-9 מחלקות, עבור ניתוחים שבוצעו בבוקר וניתוחים שבוצעו אחה"צ בנפרד. להלן גרף המציג את התוצאות (עמודות כהות מייצגות זמני ניתוח ממוצעים בשעות הבוקר, עמודות בהירות מייצגות זמני ניתוח ממוצעים בשעות אחה"צ):

Operations Time - Morning (by Hour) vs. Afternoon (by Case):



א. ישנה אסוסיאציה (association) בין המשתנה "משך ניתוח ממוצע" לבין המשתנה "זמן ביום" (המקבל את הערכים "בוקר" או "אחר הצהריים"). נכון או לא נכון? נמקו במשפט אחד (2 נק')

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ב. לאחר שראה את הגרף, סגן מנהל בית החולים הסיק שהרופאים ממהרים לסיים יותר ניתוחים כאשר הניתוחים מתבצעים בשעות אחה"צ וזאת על מנת להספיק לבצע יותר ניתוחים וכך לקבל שכר גבוה יותר. מנהל המחלקה הכירורגית טען שזה לא יתכן שהרופאים מסכנים את חיי המטופלים במודע ע"י כך שמזדרזים לקצר את הניתוחים. עם זאת, לטענתו רופאים מאריכים באופן מלאכותי את זמני הניתוח בבקרים על מנת לבצע פחות ניתוחים בבקרים וכך לעבוד פחות קשה, בעוד שהתשלום עבור זמן העבודה שלהם בבוקר קבוע. לפי הנתונים שבידיכם, מי מהם צודק (אם בכלל)? (2 נק')

_____

_____

_____

_____

ג. מנהלת בית החולים תשאלה מספר מנתחים וטענה כי ישנו משתנה מתערב (confounder) אשר מסביר את התוצאות בגרף טוב יותר מההסברים של סגן מנהל בית החולים ושל מנהל המחלקה הכירורגית. הציעו משתנה מתערב אשר סביר שגורם לתוצאות הנצפות. (3 נק')

_____

_____

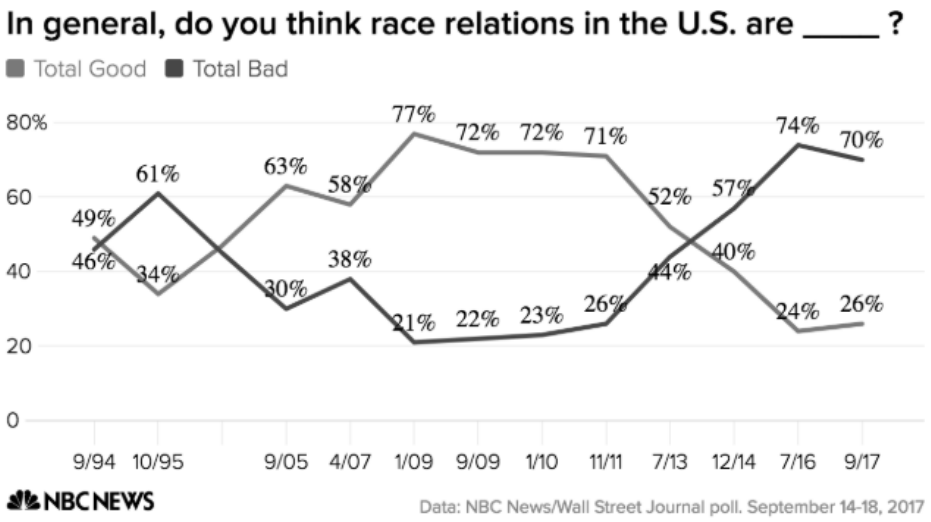
_____

_____

ת"ז: _____

<u>שאלה 8 (7 נק')</u>

נתונה הויזאוליזציה הבאה:



א. פרטו בנוגע לשתי בעיות לפחות בויזואליזציה המטעות או מבלבלות את הקורא. (הצבעים דומים רק עקב אילוצי הדפסת המבחן: צבעים דומים זו <u>אינה</u> טעות של הויזואליזציה) (4 נק')

_____

_____

_____

_____

_____

_____

ב. ציירו מחדש את הויזואליזציה כך שלא תהיה מטעה (אין צורך לדייק, רק שנבין מה הרעיון) (3 נק') [מקום לאיור]

ת"ז: _____

שאלה 9 (8 נק')

במסגרת הקורס "מבוא לניתוח נתונים", דני התבקש לייצר אלגוריתם סיווג מסוג kNN שישתמש בשני פיצ'רים (predictor_1, predictor_2) על מנת לסווג תצפיות לאחת משתי מחלקות (Class).

להלן סיכום הנתונים שדני קיבל לשם ביצוע המשימה:

```
knn_df.describe()
```

	predictor_1	predictor_2	Class
count	1372.000000	1372.000000	1372.000000
mean	0.539114	0.587301	0.444606
std	0.205003	0.219611	0.497103
min	0.000000	0.000000	0.000000
25%	0.379977	0.451451	0.000000
50%	0.543617	0.602168	0.000000
75%	0.711304	0.770363	1.000000
max	1.000000	1.000000	1.000000

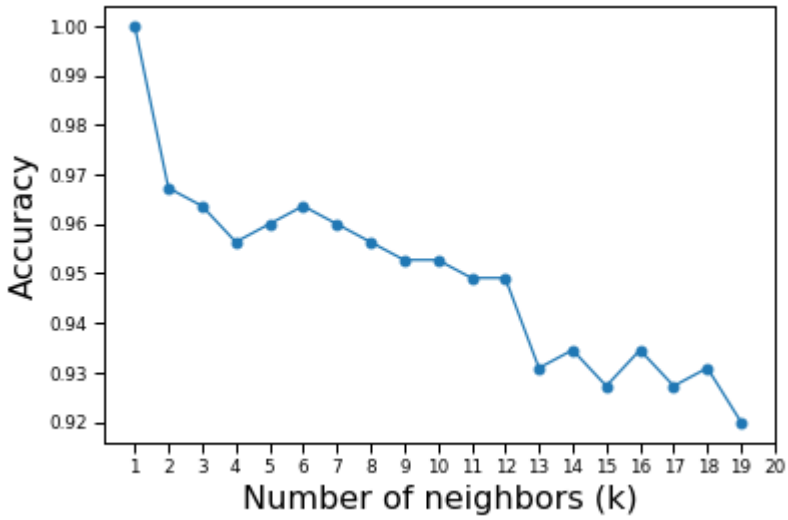
דני החליט לאמן את האלגוריתם עם ערכים שונים של k כדי לבדוק איזה מהם הוא הטוב ביותר.

להלן הקוד שדני כתב והפלט שקיבל:

```
knn_df = knn_df.sample(frac=1)

X = knn_df.loc[:, knn_df.columns != 'Class'].values
Y = knn_df.loc[:, 'Class'].values

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20)

test_scores = []
k_list = range(1, 20)
for nn in k_list:
    knn_classifier = KNeighborsClassifier(n_neighbors=nn)
    knn_classifier.fit(X, Y)
    test_scores.append(knn_classifier.score(X_test, Y_test))

# output results
plt.plot(k_list, test_scores, '-o')
plt.xlabel('Number of neighbors (k)')
plt.xticks(np.arange(1,21))
plt.ylabel('Accuracy');
```

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____



א. לפי הגרף שהתקבל (בלבד), באיזה ערך של k כדאי לדני לבחור? נמקו במשפט אחד. (2 נק')

_____

_____

ב. בקוד שדני כתב ישנה טעות מהותית שגורמת לו לקבל תוצאות שגויות. מהי מספר השורה שבה מופיעה הטעות? מהי הטעות ומדוע זו טעות? כתבו שורת קוד מתוקנת. (שימו לב: כשדני כתב את הקוד, הוא עדיין לא למד את תהליך ה-cross validation ולכן העובדה שלא השתמש בכך אינה טעות) (6 נק')

_____

_____

_____

_____

_____

_____

_____

_____

ת"ז: _____

שאלה 10 (13 נק')

תמר קראה מאמר הטוען כי 6% מבני הנוער בישראל הם טבעונים. היא חושדת כי אצלה בבית הספר אחוז הטבעונים גדול יותר. כדי לבדוק זאת, היא לקחה מדגם מקרי פשוט של 25 תלמידים בבית הספר שלה והתקבל כי 5 מהם טבעונים. לאחר מכן, תמר הריצה סימולציה שבה היא לקחה 40 מדגמים מאוכלוסייה שבה 6% טבעונים וחישבה את אחוז הטבעונים שהתקבלו בכל מדגם בסימולציה.

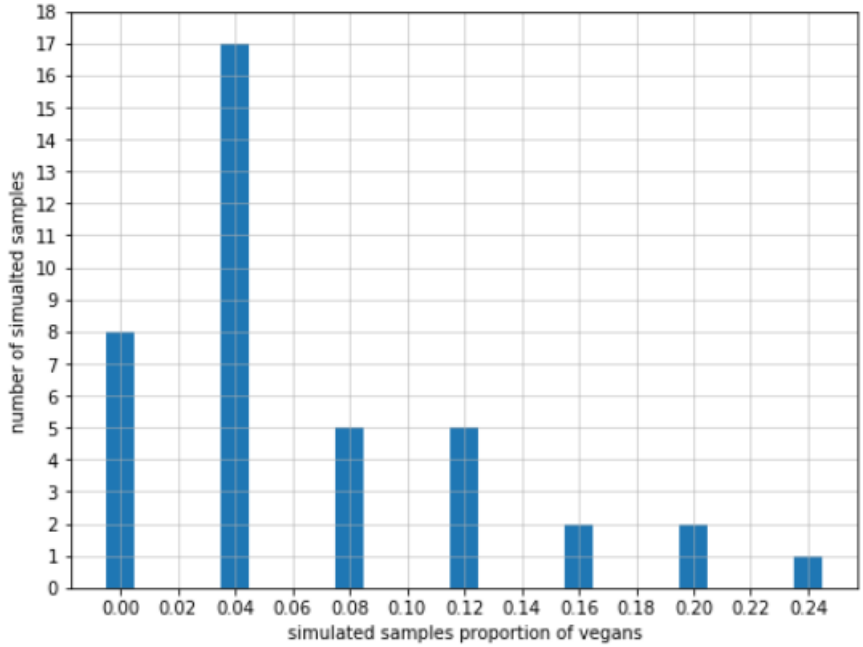
א. מהי השערת האפס אותה בודקת תמר? (2 נק')

_____

_____

ב. בהנחה שתמר ביצעה את התהליך בצורה נכונה, כמה תלמידים היו בכל מדגם אותו תמר סימלצה? (2 נק') _____

ג. להלן ההיסטוגרמה שקיבלה תמר עבור התפלגות אחוז הטבעונים תחת השערת האפס:



לפי תוצאות אלו, מהו ה-P-value של המבחן? נמקו בקצרה (5 נק')

_____

_____

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ד. ברמת מובהקות 5%, האם תמר יכולה לדחות את השערת האפס שלה? (2 נק')

_____

_____

ה. הציעו לתמר דרך פשוטה (שלא כוללת איסוף נתונים נוספים) שתגדיל את הדיוק של הסימולציה שלה. (2 נק')

_____

_____

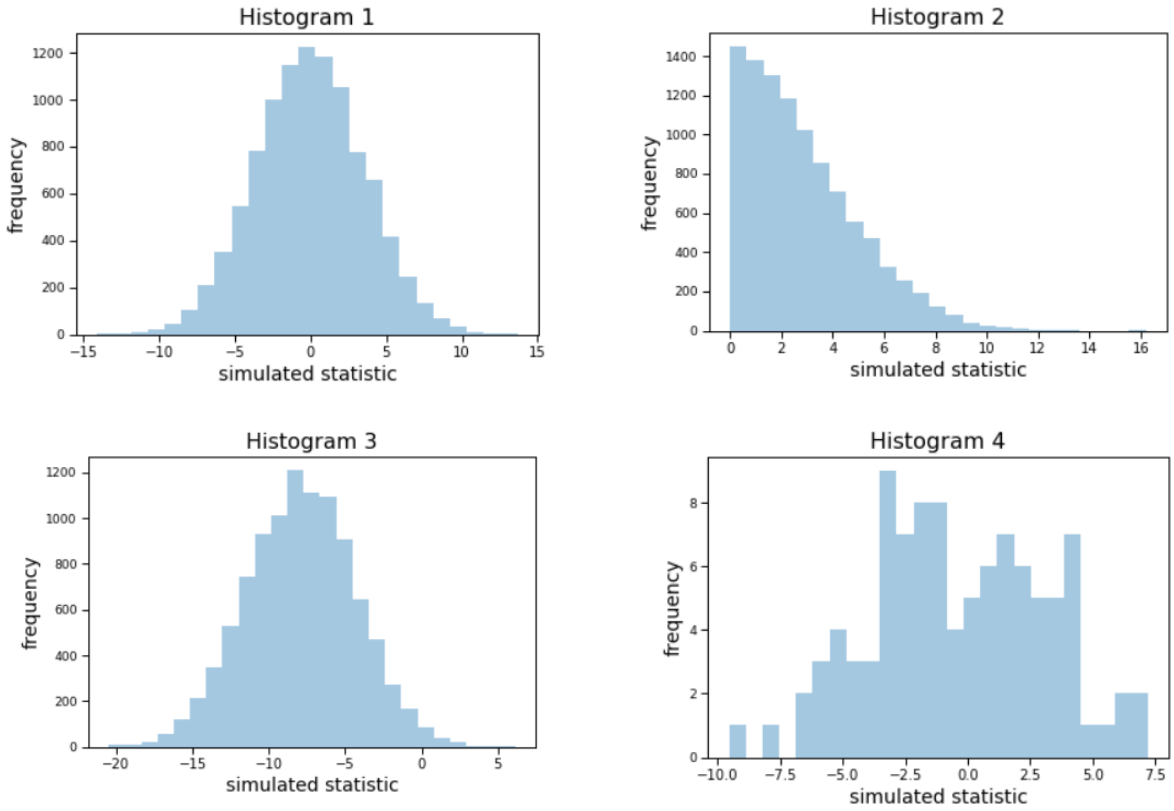
_____

ת"ז: _____

שאלה 11 (4 נק')

חוקרים רצו לבחון את האפקטיביות של חיסון השפעת השנתי. הם לקחו מדגם מקרי פשוט גדול מאוכלוסיית המתחסנים לשפעת בשנת 2017 וגילו שמתוך האנשים שנדגמו 68% לא חלו בשפעת ב-2017. בנוסף, הם לקחו מדגם מקרי פשוט גדול מאוכלוסיית המתחסנים לשפעת בשנת 2018 וגילו שמתוך האנשים שנדגמו 60% לא חלו בשפעת ב-2018. מכאן, הם חשדו כי אפקטיביות החיסון לשפעת היתה נמוכה יותר בשנת 2018 מאשר בשנת 2017. לשם בדיקת הטענה, הם הריצו 10,000 סימולציות של סטטיסטי המבחן תחת השערת האפס והציגו את ההתפלגות האמפירית של הסטטיסטי בהיסטוגרמה. אחת מארבע ההיסטוגרמות להלן היא ההיסטוגרמה אותה הם הציגו.

איזו? (הקיפו את התשובה הנכונה)



א. היסטוגרמה 1 (Histogram 1)

ב. היסטוגרמה 2 (Histogram 2)

ג. היסטוגרמה 3 (Histogram 3)

ד. היסטוגרמה 4 (Histogram 4)

ת"ז: _____

שאלה 12 (7 נק')

בידיכם נתונים אשר מסוכמים בפלט הבא:

```
df.describe(include='all')
```

	x_1	x_2	group
count	100.000000	100.000000	100
unique	NaN	NaN	2
top	NaN	NaN	B
freq	NaN	NaN	92
mean	0.070712	-0.026916	NaN
std	0.760040	0.856362	NaN
min	-2.367725	-2.297804	NaN
25%	-0.405385	-0.689449	NaN
50%	0.093605	0.022972	NaN
75%	0.621761	0.617290	NaN
max	2.100413	1.996156	NaN

מעוניינים להשתמש בנתונים אלו כדי לאמן אלגוריתם kNN שיסווג את התצפיות לקבוצות A או B (לפי משתנה group) באמצעות המשתנים x_1 ו-x_2.

א. הוחלט להשתמש בכל הנתונים כסט אימון, ללא ולידציה. כאשר יבחן דיוק האלגוריתם על נתונים חדשים הבאים מאותה אוכלוסייה (סט מבחן), מה צפוי להיות (בערך) הדיוק שיתקבל עבור k = 17? נמקו. (4 נק')

_____

_____

_____

_____

ב. הוחלט לבצע cross-validation על מנת לבחור את k. באיזה מדד ביצועים כדאי להשתמש? נמקו בקצרה. (3 נק')

_____

_____

_____