מבוא לניתוח נתונים - 096202

מבחן מועד ב' – סמסטר חורף תש"ף

תאריך הבחינה: 18.08.2020

מרצה: אורי פלונסקי

מתרגל: רפאל שללה

הוראות:

1. לרשותכם **שעה וחצי** לפתור את הבחינה.
2. מותר להשתמש במחשבון פשוט בלבד. אין להשתמש בכל חומר עזר אחר.
3. לבחינה שני חלקים. החלק ראשון כולל 6 שאלות (רוב השאלות עם מספר סעיפים) השוות יחד 77 נקודות. החלק השני כולל 3 שאלות (שוב, עם מספר סעיפים) השוות יחד 27 נקודות. (כלומר, ניתן לצבור עד 104 נקודות.)
4. ציון המבחן יהיה המקסימום מבין:
   א. מספר הנקודות שתצברו בחלק הראשון מחולק במקסימום הניקוד האפשרי (77)
   ב. מספר הנקודות שתצברו בשני החלקים יחד (לא מחולק בשום דבר)

   כלומר, החלק הראשון הוא חובה בעוד החלק השני (כולו) מהווה מגן. בפרט, אין חובה לענות על אף שאלה בחלק השני.
5. ניקוד כל שאלה (ולרוב כל סעיף) מצוין לידה.
6. במרבית השאלות והסעיפים מצוינות מגבלות על מספר השורות שיש לכתוב כתשובה. יש להקפיד לשמור על מגבלות אלו.

בהצלחה!

חלק ראשון – שאלות חובה

שאלה 1 (20 נק')

על כל אחד מהסעיפים הבאים יש לענות **האם נכון או לא נכון ולנמק** את התשובה במשפט אחד.

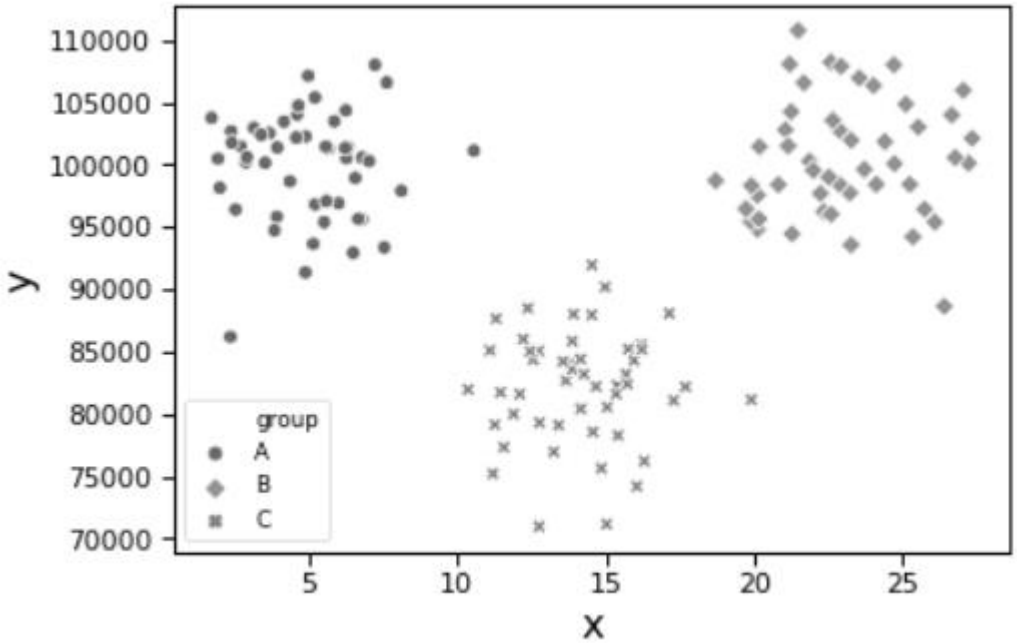
א. נתון P-value = 0.007. ההסתברות שהשערת האפס נכונה היא 0.007.

ב. כאשר משתמשים בנתוני עבר על מנת לקבל החלטות לגבי העתיד, יתכן שהטיות העבר יתעצמו.

ג. נתון כי ל-35% מהאנשים השייכים לאוכלוסייה א' ישנו גן (gene) מסוים. מדגם מקרי פשוט של אנשים השייכים לאוכלוסייה ב' נבדק ובו נמצא כי לאף אחד מהאנשים במדגם שנלקח אין את הגן הזה. מנתונים אלו ניתן להסיק, בכל רמת מובהקות סבירה, שאחוז האנשים בעלי גן זה באוכלוסייה ב' **נמוך** מאחוז האנשים בעלי גן זה באוכלוסייה א'.

ד. קו 17 אמור להגיע לתחנה כל 15 דקות בממוצע, אך שלומית חושדת שהפרש הזמנים הממוצע בין הגעת אוטובוסים עוקבים לתחנה גדול מ-15 דקות. כדי לבחון זאת, היא מגיעה 100 פעמים לתחנה בזמנים אקראיים ובכל פעם רושמת לעצמה כמה זמן היה עליה להמתין עד להגעת האוטובוס הבא. לבסוף, היא מחשבת ממוצע ל-100 זמני המתנה אלו. אם החשד של שלומית **שגוי**, והזמן הממוצע בין הגעת אוטובוסים עוקבים לתחנה הוא באמת 15 דקות, סביר להניח שהממוצע ששלומית חישבה הוא כ-15 דקות.

ה. בידינו תצפיות עם שני משתנים נומריים (x, y) ובנוסף משתנה קטגורי אחד בעל שלוש רמות (group). תרשים פיזור של הקשר בין המשתנים מוצג להלן. נניח כי לא ידוע ערכו של המשתנה הקטגורי ועל הנתונים שנותרו (תצפיות עם שני המשתנים הנומריים **בלבד**) מורץ אלגוריתם K-Means עם K=3. סביר להניח כי האלגוריתם הצליח לשחזר (כמעט לחלוטין) את הקבוצות המקוריות שמסומנות על ידי המשתנה הקטגורי. כלומר, כמעט כל הנקודות ששייכות לקבוצה A קובצו לאותו האשכול (cluster), כמעט כל הנקודות ששייכות לקבוצה B קובצו לאותו האשכול וכמעט כל הנקודות ששייכות לקבוצה C קובצו לאותו האשכול.



שאלה 2 (4 נק')

כתבה בעיתון דיווחה כי שתייה של יין יכולה למנוע מחלות לב וזאת על פי מחקר מדעי שהראה שלאנשים ששתו כוס יין מספר פעמים בשבוע היו פחות מחלות לב מאשר אלו שלא. מהי הבעיה בטיעון של הכתבה? יש להגביל את התשובה ל-3 שורות לכל היותר.

שאלה 3 (15 נק')

נתון שהערכים באוכלוסייה מסוימת הם:

{10, 10, 10, 10, 10, 20, 30, 30, 30, 30, 40, 40, 40, 50, 50}

לוקחים מדגם מקרי פשוט מתוך האוכלוסייה הזו ומקבלים: {10, 10, 20, 30, 30, 40}.

א. לכל אחד מהמדגמים הבאים, קבעו האם יתכן שהוא התקבל מתוך דגימת bootstrap.

אם תשובתכם שלילית, הסבירו במשפט אחד מדוע הדבר לא יתכן (אין צורך לנמק אם תשובתכם היא שכן יתכן): (10 נק')

1. {10, 10, 20, 30, 30, 40}
2. {10, 10, 10, 10, 40, 50}
3. {20, 20, 20, 30, 40, 40}
4. {10, 10, 10, 30, 40}

ב. לאיזה מבין המדגמים שקבעתם בסעיף א' שכן יתכן שיהיו דגימת bootstrap ישנה ההסתברות הגבוהה ביותר להופיע בדגימת bootstrap מסוימת? נמקו (5 נק')

יש להגביל התשובה ל-5 שורות לכל היותר.

שאלה 4 (4 נק')

נניח כי ברשותנו m תצפיות בסט האימון (training set) ו-n תצפיות בסט המבחן (test set). בתהליך הנקרא leave-one-out cross validation משתמשים כל פעם בתצפית אחת מתוך סט האימון כסט הוולידציה (validation set) ובכל שאר התצפיות משתמשים לשם האימון. למי מהבאים תהליך leave-one-out cross validation שקול? (העתיקו את התשובה הנכונה למחברת; אין לנמק)

א. 1-fold cross validation

ב. n-fold cross validation

ג. m-fold cross validation

ד. (m+n)-fold cross validation

שאלה 5 (19 נק')

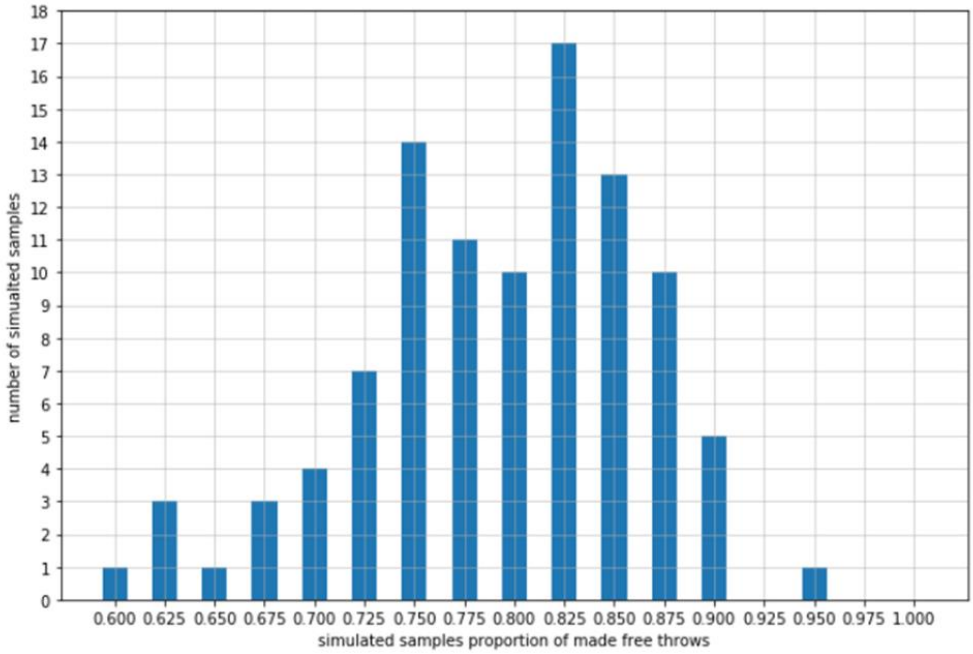
קובי אומר שהוא שחקן כדורסל מצוין. בפרט, הוא טוען שהוא קולע בממוצע 80% מזריקות העונשין שהוא זורק. עומרי חושד שקובי מגזים ושאחוז קליעות העונשין המוצלחות שלו נמוך מ-80%. כדי לבחון זאת, עומרי מבקש מקובי לזרוק 40 זריקות עונשין ברציפות. קובי מסכים ומצליח לקלוע 24 מתוכן. עומרי מעוניין לבדוק האם זוהי עדות מספיקה כדי לדחות את הטענה של קובי. לשם כך הוא מריץ סימולציה שבמסגרתה הוא לוקח 100 מדגמים שבהם 40 זריקות עונשין כל אחד.

א. מהי השערת האפס אותה בודק עומרי? (3 נק')

ב. בהנחה שעומרי ביצע את התהליך בצורה נכונה, מהו אחוז הקליעות שעומרי קבע בסימולציה שיהיו מוצלחות? (3 נק')

[השאלה ממשיכה בעמוד הבא]

ג. להלן ההיסטוגרמה שקיבל עומרי עבור התפלגות אחוז הזריקות המוצלחות תחת השערת האפס:



לפי תוצאות אלו, מהו ה-P-value של המבחן? נמקו בקצרה (7 נק')

יש להגביל את התשובה ל-4 שורות לכל היותר

ד. ברמת מובהקות 5%, האם עומרי יכול לדחות את השערת האפס שלו? הסבירו. (3 נק')

יש להגביל את התשובה ל-2 שורות לכל היותר

ה. אם עומרי היה מריץ 10,000 סימולציות במקום 100 בלבד, האם הייתם מצפים שהערך השכיח בהיסטוגרמה ישתנה? הסבירו. (3 נק')

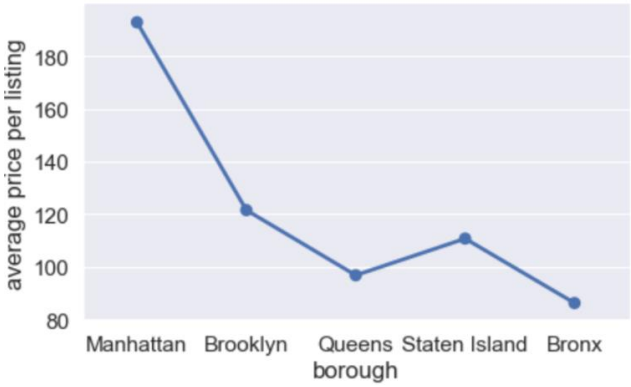
יש להגביל את התשובה ל-3 שורות לכל היותר

שאלה 6 (15 נק')

טבלת הנתונים airbnb_df כוללת מדגם מקרי פשוט של הדירות שהוצעו להשכרה בניו יורק באמצעות אתר Airbnb בשנת 2019. להלן שלוש השורות הראשונות בטבלה.

	unique_id	listing_name	host_id	host_name	area	price
0	2454	superCondo	2688	Ben	Manhattan	137
1	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	149
2	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	225

אלון ביקש להשוות בין מחיר הדירות המבוקש במנהטן (Manhattan) למחיר הדירות המבוקש בברוקלין (Brooklyn). כדי להשוות את מחירי הדירות, אלון יצר את הויזואליזציה הבאה [ראו בעמוד הבא]:



א. הסבירו מדוע זו אינה ויזואליזציה מוצלחת והציעו חלופה טובה יותר. נמקו. (ניתן, אך לא חובה, להשתמש בציור סכמטי כדי להבהיר באיזו ויזואליזציה כדאי להשתמש) (5 נק׳)
יש להגביל את התשובה ל-4 שורות לכל היותר (ו/או להוסיף ציור סכמטי)

מתוצאות הויזואליזציה, אלון חשד כי מחירי הדירות באתר Airbnb במנהטן בשנת 2019 היו גבוהים יותר מהמחירים שלהם בברוקלין. כדי לבדוק זאת, הוא תכנן לייצר רווח סמך בוטסטראפ ברמות ביטחון 95% להפרש בין המחיר הממוצע במנהטן למחיר הממוצע בברוקלין. לשם כך הוא כתב את הקוד וקיבל את הפלט הבאים.

```python
df = airbnb_df.loc[(airbnb_df['area']=="Manhattan") | (airbnb_df['area']=="Brooklyn"),:]

def diff_of_avgs(df, column_name, grouping_var):
    grpby_var = df.groupby(grouping_var)
    avgs = grpby_var[column_name].mean()
    return avgs[1] - avgs[0]

def bootstrap_mean_difference(original_sample, column_name, grouping_var, num_replications):
    sample_size = num_replications
    bstrap_mean_diffs = np.empty(num_replications)
    for i in range(num_replications):
        bootstrap_sample = original_sample.sample(sample_size, replace=True)
        resampled_mean_diff = diff_of_avgs(bootstrap_sample, column_name, grouping_var)
        bstrap_mean_diffs[i] = resampled_mean_diff

    return bstrap_mean_diffs

bstrap_diffs = bootstrap_mean_difference(df, 'price', 'area', 5000)

left = np.percentile(bstrap_diffs, 5, interpolation='higher')
right = np.percentile(bstrap_diffs, 95, interpolation='higher')
display('A 95% boostsrap confidence interval for difference between the means', [left, right])
```

```
'A 95% boostsrap confidence interval for difference between the means'

[60.6811836363122, 81.57050268237528]
```

ב. בקוד שדני כתב ישנן טעויות מהותיות שגורמות לו לקבל תוצאות שגויות. בפרט, ישנן 3 שורות בהם עשה דני טעות. מהם מספרי השורה שבהם מופיעות הטעויות? מהן הטעויות ומדוע הן טעויות? כתבו שורות קוד מתוקנות. (10 נק׳)
יש להגביל את התשובה ל-9 שורות לכל היותר.

חלק שני – שאלות מגן

שאלה 7 (9 נק')

חוקר אוסף תמונות של שלושה סוגי פרחים עם עלי כותרת אדומים: כלנית, פרג ונורית. ידוע כי התפלגות התמונות באוכלוסייה היא 45% כלניות, 30% פרגים, 25% נוריות. להלן מספר היגדים בנוגע ל-TVD (Total variation distance) בין ההתפלגות של מדגם התמונות שבידי החוקר וההתפלגות של התמונות באוכלוסייה. לכל היגד, קבעו אם הוא נכון או לא נכון ונמקו במשפט אחד:

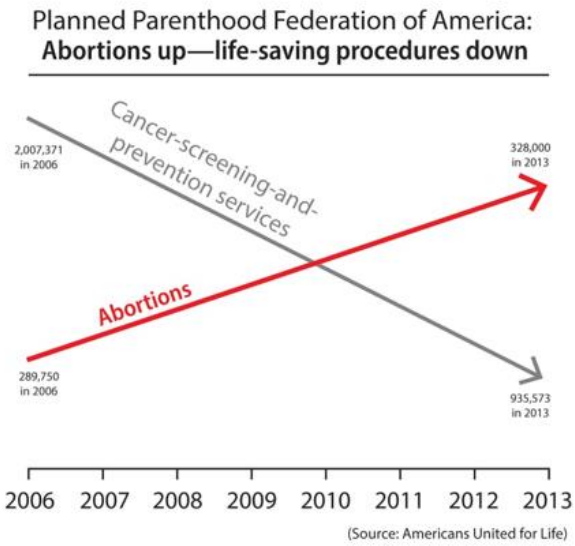
א. ה-TVD יהיה אפס כמעט תמיד

ב. ה-TVD יהיה אפס או קרוב לאפס אם בידי החוקר יש מדגם אקראי גדול של תמונות (למשל, n=1000), אבל לא בהכרח אם בידי החוקר יש מדגם קטן של תמונות (למשל n=5)

ג. אם נסמלץ את ה TVD-הרבה פעמים תחת ההנחה שההתפלגות של התמונות שבידי החוקר זהה להתפלגות באוכלוסייה, ההתפלגות האמפירית שנקבל תהיה סימטרית סביב אפס.

שאלה 8 (9 נק')

נתונה הויזאוליזציה הבאה:



א. פרטו בנוגע לשתי בעיות בויזואליזציה המטעות או מבלבלות את הקורא. (6 נק') יש להגביל את התשובה ל-6 שורות לכל היותר

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

שאלה 9 (9 נק׳)

בטבלת הנתונים technion_df יש שורה אחת לכל אחד מ-840 הסטודנטים שלומדים לתואר ראשון בפקולטה להנדסת תעשייה וניהול בטכניון בסמסטר חורף תש״פ. להלן שלוש השורות הראשונות בטבלה:

unique_id_number	gender	year_started
033952555	male	2017
345827980	female	2017
452244509	female	2018

נתון הקוד הבא:

```
sample_1 = technion_df.iloc[np.arange(1,100)]

sample_2 = technion_df.iloc[np.arange(0, technion_df.shape[0], 10)]

start = np.random.choice(np.arange(10))
sample_3 = technion_df.iloc[np.arange(start, technion_df.shape[0], 10)]
```

א. איזה מההיגדים הבאים נכון? (יתכן שיותר מאחד נכון. אין צורך לנמק) (3 נק׳)

- ☐ מדגם 1 (sample_1) הוא מדגם דטרמיניסטי (deterministic sample)
- ☐ מדגם 2 (sample_2) הוא מדגם הסתברות (probability sample)
- ☐ מדגם 3 (sample_3) הוא מדגם נוחות (convenience sample)

א. **לכל אחד משלושת המדגמים**, חשבו מהי ההסתברות שהסטודנט בעל ת״ז (unique_id_number) 033952555 נבחר להיות כלול במדגם. נמקו בקצרה כל חישוב (6 נק׳) יש להגביל את התשובה ל-7 שורות לכל היותר.