ת"ז: _____

<u>מבוא לניתוח נתונים - 094202</u>

<u>מבחן מועד א' – סמסטר אביב תשפ"א</u>

טור א'

תאריך הבחינה: 15.07.2021

מרצים: עפרה עמיר, אורי פלונסקי

מתרגלים: רפאל שללה, זהר גלעד, אלכסנדר טואיסוב

הוראות:

1. לרשותכם **שלוש שעות** לפתור את הבחינה.
2. מותר להשתמש במחשבון <u>פשוט</u> בלבד. אין להשתמש בכל חומר עזר אחר.
3. הבחינה כוללת 13 עמודים (כולל עמוד זה) ובהם 10 שאלות ושאלת בונוס. 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.** אין להוסיף מילים מעבר לשורות המיועדות. 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 אתכם לחדר הבחינה. בפרט, חובה להחזיר בסיום הבחינה את הטופס וכל מחברת בה השתמשתם.

בהצלחה!

שאלה	ניקוד
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
בונוס	
סה"כ	

ת"ז: _____

<u>שאלה 1 (15 נק')</u>

הנהלת הטכניון מעוניינת לבדוק האם במסגרת הלמידה ההיברידית היו הבדלים בהערכות המרצים (סקר ההוראה) שניתנו על ידי סטודנטים שנכחו באופן פיזי בהרצאות והערכות המרצים שניתנו על ידי סטודנטים שצפו בהרצאות באופן סינכרוני בזום (ניתן להתעלם מסטודנטים שלא נכחו בהרצאה באופן סינכרוני). הניחו שלטכניון יש גישה לסקרי ההוראה ולזהות הסטודנטים שמילאו אותם (זהות הסטודנטים מותממת, כלומר עברה אנונימיזציה). בנוסף, נאספו מכל הקורסים הלוגים (logs) של הזום (כלומר שמות הסטודנטים שהתחברו לשיעור בזום באופן סינכרוני) והוצאו מהמודל (Moodle) הרישומים של הסטודנטים להגעה פיזית לכיתה (הנתונים מהמודל גם עברו התממה ויכולים להיות מוצלבים עם נתוני סקר ההוראה).

א. נסחו את השערת האפס וההשערה האלטרנטיבית (3 נק')

_____

_____

_____

_____

ב. ציינו מהו הפרמטר שהייתם בודקים ומה יהיה סטטיסטי המבחן. (4 נק')

_____

_____

_____

ג. תארו הטיית מדידה שעשויה להיות בנתונים שנאספו, וציינו את ההשלכות האפשריות שלה על ניתוח הנתונים. (4 נק')

_____

_____

_____

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ד. הניחו שבתום הניתוח נמצא כי ישנו הבדל מובהק בין ציוני סקר ההוראה שניתנו על ידי סטודנטים שנכחו באופן פיזי לבין סטודנטים שנכחו בזום, כך שסטודנטים שנכחו באופן פיזי דירגו בממוצע את המרצים בציוני הוראה גבוהים יותר מסטודנטים שנכחו באופן מקוון. על פי נתונים אלה, הנהלת הטכניון קבעה שביחס ללמידה מקוונת, למידה בקמפוס גורמת לשביעות רצון גבוהה יותר של הסטודנטים מסגל ההוראה ומהקורס. האם לדעתכם ניתן להסיק זאת (בהנחה שהניתוח בוצע באופן סטטיסטי נכון)? אם כן, הסבירו מדוע. אם לטענתכם לא ניתן להסיק זאת, הציעו הסבר חלופי. (4 נק')

_____

_____

_____

_____

_____

_____

<u>שאלה 2 (5 נק')</u>

חוקרים מצאו כי רמת האושר **יורדת** עם העלייה בגיל. בנוסף, נמצא כי בממוצע לאנשים מבוגרים יותר יש רמת אושר **גבוהה** יותר מאשר לאנשים צעירים. ניתן להסביר פרדוקס זה בכך שהנתונים קטומים מימין (right-censored data). הסבירו למה הכוונה ב-3 משפטים לכל היותר (תוך התייחסות לקשר בין אושר וגיל).

_____

_____

_____

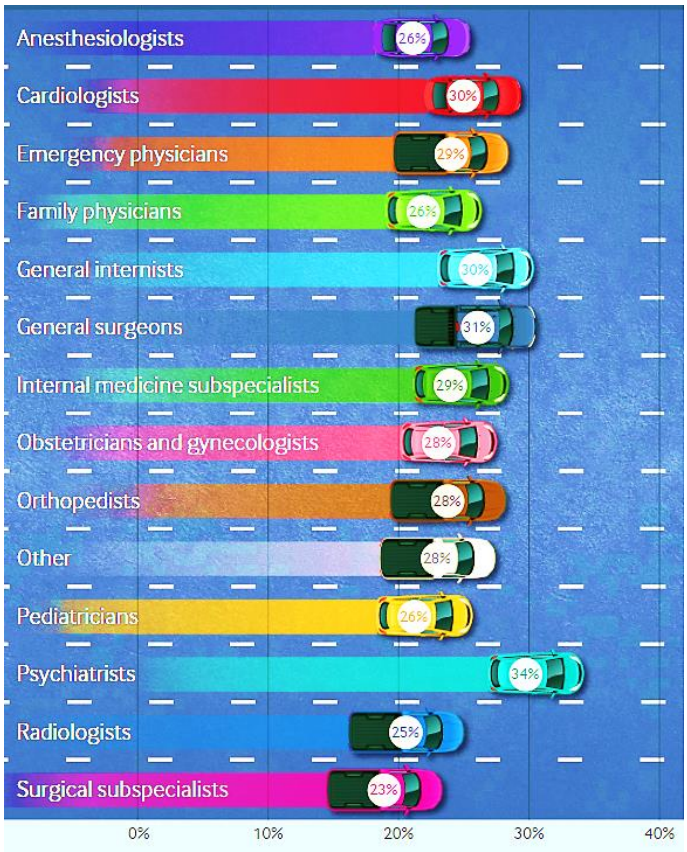
_____

_____

ת"ז: _____

<u>שאלה 3 (8 נק')</u>

הגרף הבא לקוח ממאמר אשר חקר את הנהיגה של רופאים מהתמחויות שונות, בהסתמך על בסיס נתונים של דו"חות על נהיגה מהירה בין השנים 2004-2017. "מהירות קיצונית" הוגדרה במאמר כנהיגה במהירות הגבוהה ב-20 מייל לשעה או יותר מהמהירות המותרת בכביש בו ניתן הדו"ח.
הגרף הבא מראה עבור כל תחום התמחות רפואית, את אחוז הדו"חות שניתנו על נהיגה קיצונית (מתוך כלל הדו"חות במדגם שניתנו עבר רופאים מתחום ההתמחות). למשל, מבין הדו"חות שניתנו לפסיכיאטרים (psychiatrists), 34% היו במקרים של עבירת מהירות "קיצונית".



א. ציינו בעיה מהותית הקיימת בגרף עצמו (ולא בנתונים המוצגים) והסבירו מדוע זו בעיה. (צבעים דומים מדי <u>אינה</u> בעיה של הגרף אלא אילוץ של הדפסת המבחן) (3 נק')

_____

_____

_____

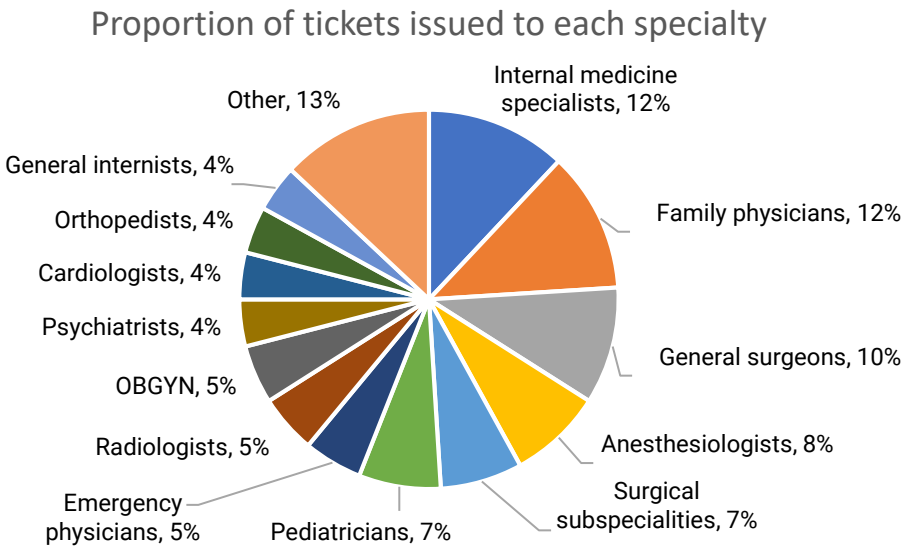
_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ב. הציעו דרך לתקן את הבעיה (הדגימו בשרטוט, לא צריך לכלול את כל הנתונים, רק הדגמה שנבין מה הרעיון) (2 נק') [מקום לאיור]

ג. הגרף הבא לקוח גם הוא מאותו המחקר, ומראה את אחוז דו"חות המהירות שחולקו לרופאים מההתמחויות השונות מתוך כלל הדו"חות שחולקו. מטרת הגרף היא להציג את הסיכוי של רופא/ה מכל אחת מההתמחויות לקבל דו"ח על עבירת מהירות ביחס לרופא/ה מהתמחויות אחרות. ציינו בעיה **בנתונים המוצגים בגרף** (ולא בויזואליזציה עצמה) אשר מקשה עלינו ללמוד על מידת הנהיגה המהירה של רופא/ה מהתמחות אחת לעומת רופא/ה מהתמחות אחרת. הציעו דרך לתקן את הנתונים המוצגים כדי להתגבר על בעיה זו. (3 נק')



_____

_____

_____

_____

ת"ז: _____

שאלה 4 (14 נק')

אתם עובדים באתר חדשותי גדול. ברשותכם נתוני גלישה באתר המכילים עבור כל כניסה לכתבה:

- מזהה הכתבה [מספר זיהוי ייחודי לכל כתבה]
- קטגוריית הכתבה [חדשות/ספורט/...]
- כתובת IP [כתובת הרשת של הגולש]
- תאריך כניסה
- שעת כניסה
- משך קריאה
- מנוי פרימיום [כן/לא]
- מזהה ייחודי למשתמש [רק עבור מנויי פרימיום]

לדוגמה שתי שורות מתוך הטבלה:

IP_address	article_id	article_category	date_enter_article	time_enter_article	time_on_article_seconds	premium_user	user_id
103.22.5.6	431587	sports	05/07/2021	10:03:22	72	True	reader123
212.13.8.12	552413	culture	05/07/2021	10:04:55	720	False	NaN

צוות ניתוח הנתונים מעוניין לאמן מסווג שיחזה אם קורא שאינו מנוי פרימיום (מזוהה על ידי כתובת IP) בחודש מסוים יירשם למינוי פרימיום בחודש שלאחר מכן.

א. הציעו 2 משתנים חדשים שתייצרו על בסיס הנתונים הגולמיים עבור כל משתמש, בהם תיעזרו בבעיית הסיווג. תארו בבירור איך יחושב המשתנה (לא צריך קוד) ומדוע אתם חושבים שהוא עשוי להיות שימושי. (6 נק')

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ב. הניחו שלאחד מהמשתנים שייצרתם יש קורלציה של 0.8− עם המשתנה של מנוי פרימיום (הניחו ש-True מיוצג על ידי הערך 1 ו-False על ידי 0). למשתנה נוסף יש קורלציה של 0.1 עם המשתנה של מנוי פרימיום. באיזה מהם הייתם מעדיפים להשתמש לצורך בעיית הסיווג? הסבירו מדוע במשפט אחד. (3 נק')

_____

_____

_____

ג. צוות ניתוח הנתונים של האתר הציע לחלק את הנתונים שנאספו בין החודשים ינואר-יוני 2021 באקראי, כך ש-80% ישמשו לאימון ו-20% לבדיקה. ציינו בעיה בחלוקה זו והציעו חלוקה חלופית לסט בדיקה וסט אימון. (5 נק')

_____

_____

_____

_____

_____

_____

<u>שאלה 5 (4 נק')</u>

תמר חישבה רווח סמך ברמת ביטחון 95% לממוצע באוכלוסייה בהתבסס על מדגם אקראי וגדול. רווח הסמך שהתקבל הוא [62.4, 71.9]. לאחר מכן אילנית קיבלה גישה לנתונים על האוכלוסייה כולה ומצאה כי הממוצע באוכלוסייה הוא 73.5. הסבירו כיצד זה יתכן (בהנחה שתמר חישבה את רווח הסמך נכון ואילנית חישבה את הממוצע נכון)

_____

_____

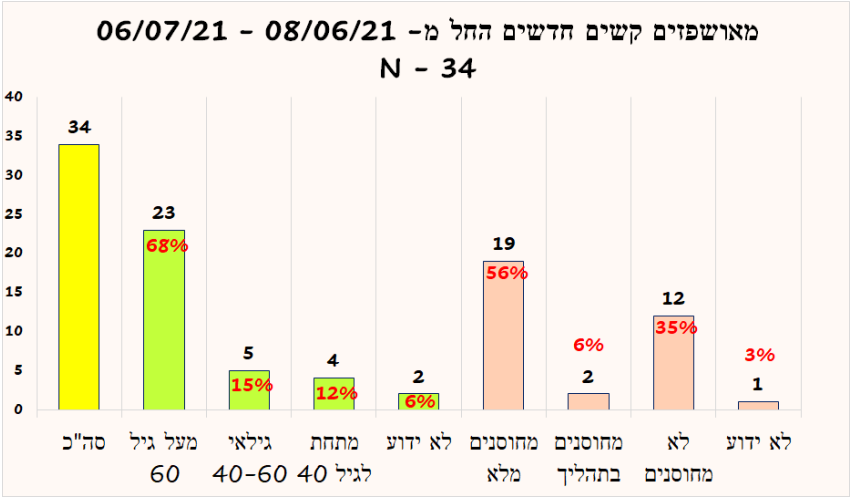
_____

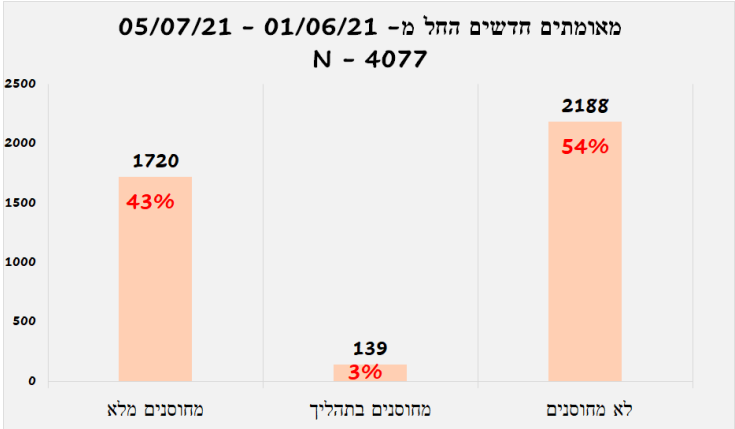
_____

ת"ז: _____

<u>שאלה 6 (8 נק')</u>

מואיז פרסם את הגרפים הבאים:





א. בתגובה, אפרת כתבה כי מאחר שיש יותר מאושפזים קשים "מחוסנים מלא" מאשר "מחוסנים בתהליך", הגרף מראה שעדיף להיות "מחוסן בתהליך" מאשר "מחוסן מלא". הסבירו מדוע לא ניתן להסיק מסקנה כזו מהגרף והציעו הסבר חלופי לנתונים. (4 נק')

_____

_____

_____

ב. יוסף כתב שמאחר ששיעור המחוסנים מלא מתוך כלל המאומתים החדשים נמוך יותר משיעור המחוסנים מלא מתוך כלל המאושפזים הקשים, הנתונים מראים שהחיסון אינו אפקטיבי במניעת מחלה קשה. הסבירו מדוע לא ניתן להסיק מסקנה זו מהגרפים והציעו הסבר חלופי. (4 נק')

_____

_____

_____

ת"ז: _____

שאלה 7 (6 נק')

בסט נתונים מסוים יש 100 תצפיות, מתוכן ל-75 יש תגית "כחול" ול-25 תגית "לבן".
אליהו חילק את הנתונים לסט אימון עם 80 תצפיות ולסט מבחן עם 20 תצפיות. הוא הריץ אלגוריתם kNN על סט האימון ומצא שה-k הטוב ביותר עבור הנתונים שלו בסט האימון הוא 11. בסט האימון, דיוק האלגוריתם של אליהו הוא 93.4% = 75/80.
כאשר אליהו בחן את דיוק האלגוריתם על סט המבחן הוא מצא להפתעתו שהדיוק שקיבל בסט המבחן הוא בדיוק 0 (אפס). כלומר, אף תצפית מסט המבחן לא סווגה נכון. הסבירו ב-4 משפטים לכל היותר כיצד זה יתכן, מה סביר שקרה לאליהו ומה הייתם ממליצים לו לעשות.

_____

_____

_____

_____

_____

_____

_____

שאלה 8 (13 נק')

הפלט של הפקודה my_df.describe() עבור סט נתונים מסוים הוא:

	variable1	variable2	variable3	target
count	3882.000000	3882.000000	3.882000e+03	3882.000000
mean	51.134982	109.883823	3.872725e+07	0.494333
std	69.851621	22.721818	4.792372e+07	0.500032
min	0.000000	34.000000	2.180000e+02	0.000000
25%	7.000000	95.000000	1.000000e+07	0.000000
50%	28.000000	106.000000	2.400000e+07	0.000000
75%	65.750000	120.000000	5.000000e+07	1.000000
max	761.000000	330.000000	7.000000e+08	1.000000

א. מהו ערך הרבעון השלישי של המשתנה variable2? (2 נק')

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

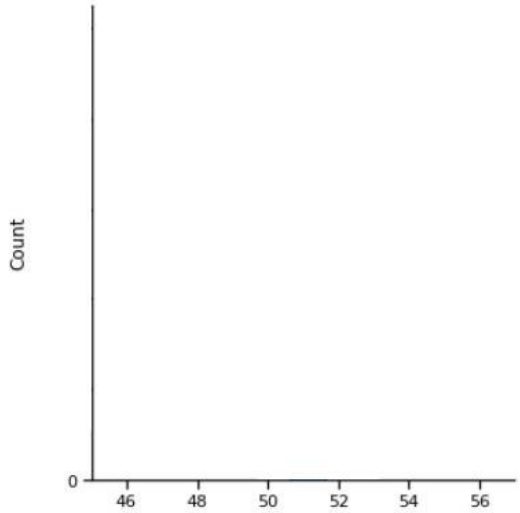
מורן הריצה את הקוד הבא על מנת לחשב רווח סמך ברמת בטחון 95% לממוצע של המשתנה variable1 באוכלוסייה:

```
def bootstrap_mean(original_sample, column_name, num_replications):
    '''This function returns an array of bootstrapped sample averages:
    original_sample: df containing the original sample
    column_name: name of column containing the variable of interest
    num_replications: number of bootstrap samples'''
    original_sample_size = original_sample.shape[0]
    original_sample_var_of_interest = original_sample[[column_name]]
    bstrap_means = np.empty(num_replications)
    for i in range(num_replications):
        bootstrap_sample = original_sample_var_of_interest.sample(original_sample_size, replace=False)
        resampled_mean = bootstrap_sample.mean()
        bstrap_means[i] = resampled_mean

    return bstrap_means

means_bootstrapped = bootstrap_mean(my_df, 'variable1', 5000)
ax=sns.displot(means_bootstrapped)
ax.set(xlim=(45,57))
```

ב. השלימו את ההיסטוגרמה שתתקבל כפלט (כלומר הפלט שיודפס למסך מקוד זה), כולל השלמת הערכים על ציר Y. רמז: ישנה טעות בקוד של מורן. (4 נק')



ג. מהו רווח הסמך שתקבל מורן כאשר תחשב אותו על סמך הקוד הנתון לעיל (כלומר הקוד שבו יש טעות)? (3 נק')

_____

ד. איתי בונה מסווג שינבא את ערכו של המשתנה target באמצעות המשתנים variable1, variable2, variable3. לשם כך חילק את הנתונים לסט אימון וסט מבחן וביצע את הליך ה-cross-validation בצורה נכונה. מהנתונים לעיל, ציינו בעיה אחת שאיתי עלול להתקל בה ללא עיבוד נוסף של הנתונים לפני הליך ה-cross-validation. (4 נק')

_____

_____

_____

ת"ז: _____

<u>שאלה 9 (12 נק')</u>

ארגון למען ילדים בסיכון החליט לייצר אלגוריתם שינבא על בסיס נתונים מדו"חות של עובדי רווחה האם הילד או הילדה בסיכון. לאחר שעובדי הארגון אימנו את המודל, הם בחנו אותו על סט מבחן של 1000 תצפיות. להלן הנתונים שהם חישבו עבור האלגוריתם:

- דיוק (accuracy) 0.89
- רגישות (sensitivity) 0.2

א. השלימו את מטריצת הבלבול (8 נק'):

	תגית אמיתית: בסיכון	תגית אמיתית: לא בסיכון	סה"כ
ניבוי המודל: בסיכון			50
ניבוי המודל: לא בסיכון			950
סה"כ	100	900	1000

ב. האם הייתם ממליצים לארגון להשתמש במודל זה כדי לנבא האם ילדים חדשים (שאין עליהם נתונים) נמצאים בסיכון? הסבירו ב-3 משפטים לכל היותר. (4 נק')

_____

_____

_____

_____

_____

ת"ז: _____

<u>שאלה 10 (15 נק')</u>

ענו נכון או לא נכון. נמקו תשובתכם במשפט אחד.

א. ההנחה העומדת בבסיס שיטת bootstrap היא שההתפלגות של המדגם מייצגת בקירוב טוב את ההתפלגות של הסטטיסטי תחת השערת האפס. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ב. אם הדיוק (accuracy) של מסווג א' גבוה מהדיוק (accuracy) של מסווג ב', ה-true positive rate (TPR) של מסווג א' בהכרח גבוה מה-true positive rate (TPR) של מסווג ב'. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ג. כשמפעילים על סט נתונים מסוים 5-fold cross validation, משתמשים בכל אחת מהתצפיות לאימון המודלים הנבחנים בדיוק 4 פעמים. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ד. בגרף קו שמייצג שינוי, התחלת ציר ה-Y מאפס לא הכרחית ואף עלולה להטעות.
נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ה. בידינו נתונים על כל ציוני המבחן במבוא לניתוח נתונים בטכניון בסמסטר אביב תשפ"א. דרך אחת סבירה להעריך את הממוצע במבחן היא לקחת דגימת בוטסטראפ מתוך נתונים אלו ולחשב לממוצע רווח סמך ברמת ביטחון 95%. נכון / לא נכון.

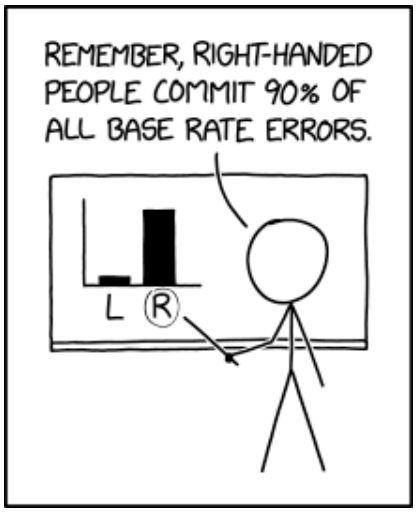
נימוק (במשפט אחד): _____

_____

ת"ז: _____

<u>שאלת בונוס (2 נק')</u>

הסבירו בקצרה מדוע קריקטורה זו (אמורה להיות) מצחיקה. השתמשו במושגים שנלמדו בקורס:



_____

_____

_____

_____

_____

_____

_____