עבודה מסכמת ב-R

מבוא לסטטיסטיקה, 2021-2022

**הנחיות כלליות**

יש להגיש את העבודה בזוגות עד לתאריך **20.3.2022 ב-23:59**. ההגשה תתבצע לתיבת הגשה ייעודית במודל, אליה יש להגיש שלושה קבצים:

1. **קובץ קוד** (R Script).
2. **קובץ Word**, שיכלול הסברים מפורטים לגבי סעיפי העבודה, יחד עם תוצאות וגרפים מה-R.
3. **קובץ נתונים**, שבו נעשה שימוש לצורך העבודה.

שמות כל הקבצים יהיו מספרי תעודות הזהות של המגישים/ות (למשל: 123456789\_987654321). **רק נציג/ה אחד/ת מתבקש/ת להגיש את הקבצים לתיבת ההגשה במודל**.

שימו לב, מטרת העבודה היא לבדוק כיצד אתם/ן מתמודדים/ות עם למידה עצמאית של R, בהתבסס על השיעורים והתרגולים. לפיכך, צוות הקורס לא יענה על שאלות הקשורות ישירות בעבודה. בכל מקרה של התלבטות או ספק, החליטו עצמאית כיצד לפעול ונמקו זאת תחת הסעיף המתאים בקובץ ה-Word.

**1.1. מניעת הונאה אקדמית בהגשת עבודות**

במטרה למנוע העתקה, מרמה, ו/או מצב שבו רק אחד/ת מהמגישים/ות ביצע/ה את העבודה, אנו נדגום באופן אקראי סטודנטים/ות שיידרשו להגן על העבודה (להגן = שיחה עם מתרגל/ת, שבה יישאלו שאלות על העבודה). מטרת שיחות אלה היא לוודא שאתם/ן כתבתם/ן את העבודה באופן עצמאי.

אם יתעורר חשד לאחד מהמצבים שתואר לעיל, הדבר ייחשב לעבירת משמעת מצד הזוג, גם אם רק אחד/ת מהם/ן ביצע/ה את העבירה בפועל. עבירות אלה יטופלו בחומרה על ידי וועדת המשמעת האוניברסיטאית.

**2. מקורות למציאת נתונים**

עליכם/ן לחפש קובץ נתונים שמעניין אתכם/ן וכולל מעל ל-5 משתנים, מתוכם לפחות שניים קטגוריאליים (בסולם שמי או סדר). כמו כן, הקובץ צריך להכיל לפחות 50 תצפיות (שורות). ניתן לחפש קובץ נתונים באופן עצמאי או במקורות המצורפים להלן:

1. קבצי נתונים הנמצאים בחבילות שונות ב-R – [לחצו כאן](https://vincentarelbundock.github.io/Rdatasets/datasets.html).
2. [Kaggle](https://www.kaggle.com/datasets).
3. [Google Dataset Search](https://datasetsearch.research.google.com/).

שימו לב – עליכם/ן לצרף את הקובץ לעבודה, לכתוב את מקורו ולצרף במסגרת מסמך ה-Word תיאור קצר של המשתנים בקובץ.

**על כל זוג לעבוד על קובץ נתונים אחר** – לשם כך, עליכם/ן לבדוק שהקובץ שברצונכם/ן לעבוד עליו לא נתפס ע"י זוג אחר ב[**טבלה זו**](https://docs.google.com/spreadsheets/d/18wnbsEwT_3bXO5NAeRFrxoVOaF7n_MOJB-SVfMRph-I/edit?usp=sharing). אם הקובץ פנוי, יש למלא את [**הטופס המקוון**](https://forms.gle/mHiLN7WdLiCUvJ9G9)על מנת לשריין את קובץ הנתונים.

**3. אופן כתיבת הקוד**

הקפידו על כתיבת קוד מסודרת – כותרת לכל סעיף, בשילוב הערות, שמבהירות לכם/ן ולנו את החשיבה שהובילה לכתיבת הקוד (להוספת הערות – השתמשו ב-**#**). בנוסף, יש לוודא שהקוד רץ במלואו ללא שגיאות.

אם תוך כדי העבודה תצוץ הודעת שאינכם/ן מצליחים/ות לפתור, חפשו אותה ב-Google. רוב הסיכויים שתמצאו את הפתרון שם.

סעיפי העבודה מנוסחים בלשון נקבה אך פונים לשני המינים.

**סעיפי העבודה**

**א.** בחרו שני משתנים מספריים מקובץ הנתונים שלכן ובצעו ביניהם מכפלה וקטורית ליצירת משתנה חדש. כעת צרו לולאה שתרוץ 10000 פעמים. בכל פעם הלולאה תדגום מקרית 100 תצפיות מתוך המשתנה החדש (בדגימה עם החזרה), תחשב את הטווח הבין רבעוני של 100 התצפיות ותשמור את הערך במשתנה חדש. הציגו בתרשים את התפלגות המשתנה החדש בעל 10000 הערכים ותארו במילים כיצד היא נראית. כיצד תשתנה ההתפלגות אילו בכל פעם תדגמו 20 תצפיות במקום 100? מדוע?

**ב**. חשבו את מדד הפיזור AD (Absolute Deviation - ממוצע הסטיות המוחלטות מהחציון) על משתנה בקובץ הנתונים שבחרתן. שימרו את תוצאת הAD של המשתנה בשם *ad*. כעת, דיגמו 100 פעמים חמישים ערכים מתוך המשתנה שבחרתן, חשבו את AD עבור כל מדגם כזה, והכניסו את היחס בין הAD המקורי (החישוב של AD עבור כל הערכים של המשתנה המקורי) לAD של המדגם החדש לוקטור בשם *ad\_ratio*. חשבו את השונות של הוקטור *ad\_ratio*, מה המשמעות של השונות הזו?

**ג**. בחרו שני משתנים מספריים מקובץ הנתונים, חשבו את המתאם ביניהם והכניסו אותו למשתנה בשם . כעת, כתבו לולאה שתחשב את המתאם 250 פעמים על 70% מהתצפיות במשתנה, שיידגמו מקרית בכל פעם. הכניסו את התוצאות לווקטור בשם . השתמשו בחבילת *ggplot2* להצגת התפלגות התוצאות בווקטור *cor\_vector2*, וסמנו על תרשים ההתפלגות את הערך *cor\_value1*.

**ד**. צרו משתנה מקרי בעל 1000 ערכים הנלקחים מתוך התפלגות נורמלית. צרו לולאה הדוגמת בכל פעם 20% מהערכים מתוך המשתנה שעוצרת רק אם **הפער** בין הממוצע, לחציון של הערכים שדגמתן גדול מהערך 0.01 (למשל, חציון=22, ממוצע =21). כמה פעמים רצה הלולאה כדי לעמוד בתנאי? מה מסקנתכם מכך?

**ה**. בחרו שני משתנים קטגוריאליים (=בסולם שמי) מהקובץ. חשבו את ההסתברות המותנית עבור כל אחת מהקטגוריות של המשתנה הראשון בכל אחת מהקטגוריות של המשתנה השני. כיתבו את ההסתברויות המותנות ואת הסתברויות החיתוכים. האם יש תלות בין המשתנים?

**ו**. המציאו שאלה בקומבינטוריקה ופתרו אותן בהתבסס על הנלמד בנושא באמצעות פרמוטציה ו/או קומבינציה .

**ז**. בחרו מתוך קובץ הנתונים משתנה רציף וחשבו עבורו את אחוז הסטיות מהחציון, אחוז הסטיות מהממוצע ואחוז הסטיות מהשכיח. בדקו לאיזה מדד מרכז יש את הערך המקסימלי ולאיזה הערך המינימאלי מבין השלושה. מדוע זה כך?

**ח**. חישבו על שאלה מעניינת העוסקת במשתנים הקיימים בקובץ, ובצעו ניתוח (או ניתוחים) הבודקים שאלה זו. כתבו את מסקנתכם מהניתוח בקובץ ה-word.

**ט**. הציגו שתי פונקציות שלא נלמדו בקורס והסבירו את השימוש בהן. כיתבו כיצד "גיליתן" את הפונקציות האלה, והשתמשו בהם על קובץ הנתונים. ניתן להשתמש בפונקציות אלה לשם ניתוח קובץ הנתונים, הצגה גרפית או לשם פתרון שאלה כלשהי.

בהצלחה!