עבודה הגשה בהסקה סטטיסטית סמסטר ב׳ תשפ״ב 2022

הנחיות כלליות

את העבודה יש להגיש **בזוגות** עד לתאריך 15.8.22 בשעה 8:00 בבוקר; ההגשה תתבצע במודל. יש להגיש שלושה קבצים: קובץ ה-script ב-R, קובץ word, ואת קובץ הנתונים שבו השתמשתם (אפשרי להגיש גם קובץ project). שמות כל הקבצים יהיו מספרי ת.ז של המגישים (למשל: 987654321_012345678). שימו לב – רק אחד משני הסטודנטים יעלה למודל את הקבצים.

כדי למנוע העתקות, מרמה, ו/או מצב שבו רק אחד/ת מבין זוג הסטודנטים שמגיש את העבודה אכן ביצע אותה, אנו נדגום באופן אקראי סטודנטים שיידרשו להגן על עבודתם בפגישה עימנו בזום (להגן = פגישה עם אחד/ת מחברי הסגל של הקורס, שבה הסטודנט נשאל שאלות על העבודה). מטרתו של מפגש זה לוודא שאכן הסטודנט ביצע את העבודה בעצמו/ה. אם יתעורר חשד להעתקה, או לכך שהסטודנט/ית לא השתתפ/ה בפועל בעבודה, הדבר ייחשב לעבירת משמעת מצד שני חברי הצוות, אשר תטופל בהתאם.

נדגיש כי מידע שלא יופיע במסמך הוורד לא ייבדק. על העבודה להיכתב בפונט David, פונט Times new roman לאנגלית, גודל 12, רווח כפול ושוליים רגילים, מיושר לימין. במידה ואתם בוחרים לכתוב כותרות וכותרות משנה, יש לכתוב אותם לפי כללי APA-7, במבנה הבא:

Centered, Bold, Title Case Heading	**כותרת רמה 1, במרכז, מעובה**
Flush Left, Bold, Title Case Heading Begin body text indented, here…...	**כותרת רמה 2, בצד ימין, מעובה** טקסט מתחיל מתחת לכותרת בפסקה מוזחת פנימה...
Flush Left, Bold Italic, Title Case Heading Begin body text indented, here…...	***כותרת רמה 3, בצד ימין, מעובה נטוי*** טקסט מתחיל מתחת לכותרת בפסקה מוזחת פנימה...
Indented, Bold, Title Case Heading, Ending With a Period. Begin text after period…	**כותרת רמה 4, בצד ימין, מוזחת פנימה, מעובה.** הטקסט מתחיל מיד לאחר הנקודה...
Indented, Bold Italic, Title Case Heading, Ending With a Period. Begin body text after the period.	***כותרת רמה 5, בצד ימין, מוזחת פנימה, מעובה נטוי.*** הטקסט מתחיל מיד לאחר הנקודה...

ראו: https://apastyle.apa.org/style-grammar-guidelines/paper-format/headings

אי עמידה בכללי ההגשה עלולה לגרור הורדת ניקוד של עד 5 נקודות, לשיקול בודק המטלה.

קובץ נתונים

עליכם למצוא קובץ נתונים אשר עומד בהנחיות המופיעות בהמשך. ניתן להשתמש בקבצי הנתונים הקיימים בחבילות הבסיס של R. כמו כן, האתרים הבאים מכילים קבצי נתונים רבים, ו/או מאפשרים חיפוש של קבצי נתונים:

https://www.kaggle.com/datasets

https://datasetsearch.research.google.com/

http://mlr.cs.umass.edu/ml/

בחרו קובץ אחד שהנתונים שהוא מכיל מעניינים אתכם, וכולל 5 משתנים או יותר, מתוכם לפחות שניים הם קטגוריאליים (בסולם שמי או סדר) ולפחות שניים הם רציפים (בסולם רווח או מנה). במידת הצורך, תוכלו להפוך בעצמכם משתנה רציף לקטגוריאלי. עליכם לצרף את הקובץ לעבודה, לכתוב את מקורו, ולצרף תיאור קצר של המשתנים בתחילת קובץ ה-word אותו אתם/ן מגישים.

על כל זוג לעבוד על קובץ נתונים אחר; בכדי לוודא זאת, עליכם לעבור על **המסמך הזה** ולוודא כי קובץ הנתונים עליו אתם רוצים לעבוד אינו בשימוש על-ידי זוג אחר, ולאחר מכן למלא את תעודות הזהות של המגישים, שם הקובץ שהשתמשתם בו והמקור שממנו נלקח **בטופס הזה**. לפני שאתם מתחילים לעבוד מחובתכם לוודא שקובץ הנתונים שלכם לא נמצא כבר בשימוש על ידי זוג אחר.

המלצות לאופן כתיבת הקוד:

הקפידו על כתיבת הערות בקוד שמבהירות לכם ולבודק העבודה את החשיבה שהובילה לבחירת התוכן של הקוד; כל הערה אותה אתם כותבים, התחילו בסימון של סולמית. במידה ואתם נתקלים בהודעת שגיאה שאינכם מצליחים לפתור, פנו ל-Google, רוב הסיכויים שתמצאו שם את הפתרון. אם החיפוש לא עזר, **המדריך הזה** כולל מספר רב של משאבים מקוונים והסברים כיצד להשתמש בהם.

כל המידע הנדרש לכתיבת העבודה מופיע בקובץ זה; אין לשלוח שאלות בצורה פרטנית בנוגע לעבודה. רק הועד יכול לשלוח שאלות בצורה מרוכזת, עד 2 פעמים לפני מועד הגשת העבודה – המועדים לבחירת הועד.

עבודה מסכמת ב-R

1. עליכם לחשוב על 3 שאלות מעניינות שניתן לענות עליהן באמצעות הנתונים בקובץ, לכתוב מה הן השאלות, ולבצע את הניתוח הסטטיסטי המתאים לכל שאלה כזו. על הניתוח של כל אחת מהשאלות לכלול: (1) סטטיסטיקה תיאורית (מדדים שונים שמתארים את המשתנה; ממוצע, סטיות תקן, גודל המדגם, חלוקה לפי מגדר וכו'), ו-(2) סטטיסטיקה היסקית (מבחן סטטיסטי או אמידה). המענה על כל אחת מהשאלות צריך להיות באמצעות פרוצדורה סטטיסטית שונה. במידת הצורך, תוכלו ליצור משתנים חדשים בקובץ, על סמך הנתונים הקיימים.

2. קובץ הוורד צריך לכלול, עבור כל שאלה, את הסעיפים הבאים:

- שאלת המחקר

- הסבר לגבי הניתוח שבו בחרתם; הסבירו בשניים-שלושה משפטים מדוע זהו הניתוח המתאים.

- בדיקת ההנחות הדרושות; כל הפרוצדורות בהן נעשה שימוש לבדיקת ההנחות צריכות להופיע במסמך הוורד. לשם התרגיל, עליכם לבצע את הניתוח גם אם ההנחות אינן מתקיימות ולציין בגוף העבודה שההנחות לא התקיימו.

- תיאור קצר (עד 3 שורות) של התוצאות; תיאור התוצאות ברמה הסטטיסטית צריך לכלול לכל הפחות את סטטיסטי המבחן, ד"ח (במידה ויש), p-value. לדוגמה, במבחן t לרוב נראה את הדיווח הסטטיסטי הבא:

$t(\text{df}) = \text{t value}, p = \text{p value}$

- המסקנה הסופית לגבי השאלה ששאלתם.

- יש להוסיף צילום פלט של התוצאה שקיבלתם ב-R

סך כל אורך תשובה בודדת הינו 8-10 שורות.

שימו לב, מידע שלא יופיע במסמך הוורד לא יקבל נקודות. אנו רוצים לראות שאתם יודעים להפיק את המידע הרלוונטי מהפלט, ולהבין כיצד המידע עונה על השאלה שאתם חוקרים.

4. עבור אחת מהשאלות אותן שאלתם/ן עליכם/ן ליצור גרף באמצעות ggplot2, ולצרף צילום רלוונטי בקובץ התשובות, במקום המתאים; נקודות בונוס יינתנו עבור גרפים מושקעים במיוחד.

5. עבור אחת מהשאלות אותן שאלתם/ן עליכם/ן לחשב באמצעות תוכנת G*Power עוצמה סטטיסטית וציינו מהו גודל המדגם הדרוש (קרי, עוצמת מבחן אפריורית). כפי שנלמד, חשבו גודל מדגם בעבור עוצמה סטטיסטית של 80% במבחן דו-זנבי, עבור אפקט בגודל בינוני, ואלפא של 5%. צרפו צילום מסך רלוונטי מתוך התוכנה.

6. עבור אחת מהשאלות אותן שאלתם/ן עליכם/ן לחשב Bayes Factor באמצעות תוכנת JASP או JAMOVI. הסבירו את משמעות התוצאה שהתקבלה, והוסיפו צילום של הפלט הרלוונטי מתוך התוכנה.

7. עבור אחת מהשאלות אותן שאלתם/ן עליכם/ן לחשב מדד גודל אפקט רלוונטי.

בהצלחה!

פסקה לדוגמה

(שאלת המחקר)

שיערנו שאחוז המצביעים לדונלד טראמפ במדינת נברסקה בבחירות לנשיאות ב-2020 היה שונה מאחוז המצביעים לטראמפ בשאר ארה"ב בשנה זו ; ברמת מובהקות של 5%.

(הסבר לגבי הניתוח)

כפי שנלמד בקורס, המבחן הסטטיסטי המתאים הוא מבחן Z דו-זנבי לפרופורציה. זאת מכיוון שהפרמטר הוא פרופורציית המצביעים מתוך סך המצביעים.

(בדיקת הנחות)

<u>הנחות :</u>

- דגימה מקרית

- קירוב בינומי לנורמלי : $p = 0.46, N = 40, N*p > 5$. הנחת הקירוב הבינומי לנורמלי מתקיימת.

(תוצאות)

$$\hat{p} = 0.6, \qquad Z_{\hat{p}} = 1.39, \qquad p = .082$$

התקבלה תוצאה לא מובהקת ($p = .082$), לכן נקבל את השערת האפס.

(מסקנה)

נסיק כי אחוז המצביעים לטראמפ במדינת נברסקה בבחירות לנשיאות ב-2020 לא היה שונה באופן מובהק מאחוז המצביעים לטארמפ בשאר ארה"ב בשנה זו.

(פלט ב-R)

```
        Exact binomial test

data:  sum(elect_sample) and 40
number of successes = 24, number of trials = 40, p-value = 0.08216
alternative hypothesis: true probability of success is not equal to 0.46
95 percent confidence interval:
 0.4332671 0.7513500
sample estimates:
probability of success
```