

## פרק עשירי

## השערת המחקר ובדיקתה

הכלים הכמותיים למדידת מפגשים מאפשרים הגדרת "מפגשים מכונסים" באופן מסויים, ובכך מאפשרים הצגה בהירה יותר ומדוייקת של השערת המחקר לעומת השערת האפס:

## ידוע כי:

מפגשים מכונסים של צפני ELS קשורים מושגית על פני טבלאות דו-ממדיות עשויים להתרחש באקראי גם בטקסטים אחרים. שכיחות המפגשים תתאים למודל האקראי.

## השערת האפס

מפגשים מכונסים של צפני ELS קשורים מושגית על פני טבלאות דו-ממדיות מתרחשים גם בספר בראשית כצפוי לפי המודל האקראי.

## השערת המחקר

מפגשים מכונסים של צפני ELS קשורים מושגית על פני טבלאות דו-ממדיות מתרחשים בספר בראשית יותר מן הצפוי לפי המודל האקראי.

## בדיקת השערת המחקר

כדי לבדוק את השערת המחקר, ננקט מהלך בן ארבעה שלבים:

#### שלב א:

מגדירים את "מידת הקרבה המכילת" בין שני ביטויים. זה מספר בין 0 ל-1. ככל שמופעי הביטויים כצפני ELS מינימליים בקטעים גדולים יותר, וככל שהם קרובים יותר זה לזה ופחות מפוזרים על פני הטבלאות, אזי "מידת הקרבה המכילת" קטנה וקרובה יותר ל-0.

#### שלב ב:

מגדירים שיטת סיכום של ערכי "מידת הקרבה" עבור קבוצה של זוגות ביטויים. הסיכום נעשה באמצעות סטטיסטי (פונקציה) המודד את "הנטייה הכוללת לקרבה". הסטטיסטי מקבל ערך בין 0 ל-1. אם בקבוצה יש הרבה זוגות שלהם "מידת קרבה מכילת" קטנה – הסטטיסטי יקבל ערך קרוב יותר ל-0.

#### שלב ג:

בוחרים רשימת זוגות מלים קשורות מושגית לבדיקה, ומחשבים את המידות שהוגדרו בשלבים א-ב.

#### שלב ד:

בודקים במבחן מובהקות האם הערך של הסטטיסטי עבור רשימת זוגות המלים קטן ממה שצפוי לקרות באקראי, ומה הסיכוי לקבל ערך כל כך קטן.

#### שלב א מתואר לעיל בפרק על כימות המפגשים.

שלב ב בחינת השערת המחקר נעשית באמצעות מדגם של זוגות ביטויים. הכוונה במונח "מדגם" היא, כי באמצעות קבוצה כזאת של זוגות ביטויים אנו מקווים לדגום את הטקסט הנסתר. בניית מדגמים היא משימה של שלב ג. כאן תידון השאלה הבאה: בהינתן מדגם, כיצד יש להעריך את הצלחת המדגם?

"מידת הקרבה המכילת" מודדת את הסיכוי למפגש עבור זוג ביטויים בודד. באמצעותה אפשר לקבל קבוצה של מספרים עבור קבוצה של זוגות. אך כיצד נעריך מה הסיכוי לקבל קבוצה "כזאת" של מספרים?

דרוש אפוא איזשהו מספר מסכם, אשר ייתן את ההסתברות למדגם כולו. בעצם, זו "מידה כוללת לקרבה" עבור כל הצמדים במדגם. בנספח 3 הוגדרו שתי שיטות סיכום של ערכי "מידת הקרבה המכילת" עבור קבוצה של זוגות ביטויים. בכל שיטה, הסיכום נעשה באמצעות סטטיסטי (פונקציה) המודד את מידת "הנטייה הכוללת לקרבה". הסטטיסטי מקבל ערך בין 0 ל-1. אם

בקבוצה יש הרבה זוגות שלהם "מידת קרבה מכוילת" קטנה – הסטטיסטי יקבל ערך קרוב יותר ל-

0. שתי המידות (הסטטיסטים) שהוגדרו בנספח א3 יסומנו  $P_1$  ו-  $P_2$ .

הגדרה מתמטית מלאה של שני הסטטיסטים נמצאת בנספח הנ"ל. כאן רק נעיר כי כל אחת

משתי המידות משמשת מעין "גלאי" להצפנה המשוערת.

•  $P_1$ : לפי מידה זו מונים את מספר הזוגות (מכלל הזוגות בקבוצה) עבורן התקבלה "מידת

הקרבה המכוילת" ב"אזור ההצלחה", אשר הוגדר א-פריורי כמרווח בין 0 ל- 0.2,

ומחשבים מה הסיכוי לקבל באקראי את הערך המתקבל.

•  $P_2$ : מידה זו נבנתה כך, שהיא רגישה לגודלם של ערכי "מידת הקרבה המכוילת" עבור כל

הזוגות בקבוצה.

לכל סטטיסטי,  $P_1$  או  $P_2$ , יש יתרונות וחסרונות.

(א) למשל, ל-  $P_1$  יש יתרון, שאין השפעה של "גודל הכשלון" של הזוגות "הנכשלים". אנחנו

מחפשים הצפנה – זוגות "מוצפנים". לא מעניין אותנו כיצד בדיוק מתנהגים אותם זוגות שאינם

מוצפנים. עניין זה עצמו הוא חסרון של  $P_2$ , הרגישה לתוצאות באזור הכשלון (סמוך ל-1).

(ב) לעומת זאת ל-  $P_2$  יש יתרון, שהיא רגישה לחדות ההצלחה. עניין זה עצמו חסר ל-  $P_1$ ,

שעבורה כל תוצאה ב"אזור ההצלחה" – ערכה שווה.

**שלב ג** מתואר בשער הבא, השער השני.