ב"ה

מדידת מִדגמי "כותרת"

מאת דורון ויצטום

כאן נטפל במדגמים א-סימטריים מטיפוס S, שנכנה אותם בשם מדגמי "כותרת". בסעיף א' נגדיר מדגמים אלה ונציג דוגמאות. בסעיף ב' נגדיר את המקרה הכללי של מבחן הראנדומיזציה עבור מדגמים אלה, ובסעיף ג' תוצג דוגמא מוחשית ליישום מבחן הראנדומיזציה.

א. מִדגמים מטיפוס S (מדגמי "כותרת")

מדגם מטיפוס S הוא מדגם א-סימטרי הניתן להצגה כקבוצת זוגות הביטויים

$$(w,w_1),(w,w_2),\ldots,(w,w_n)$$

בהם הביטוי w ("הכותרת") משותף לכל הזוגות. המדגם מהווה בסיס לבדיקת מפגשים – בין "הכותרת", לכל אחד מן הביטויים האחרים המופיעים כמד"שים מינימליים. ביטוי הכותרת יוצג בשני אופנים יסודיים:

1. באמצעות המד"שים של הביטוי.
2. באמצעות הופעות הביטוי בטקסט כרצף אותיות ("דילוג" של 1 קדימה או אחורנית).

שים לב, לא בכל ביטוי כותרת קיימים שני האופנים (לעתים הביטוי אינו מופיע כרצף אותיות בטקסט, ולעתים – אינו מופיע כמד"שים). כל אחד משני האופנים הנ"ל מתפצל לשתי אפשרויות:

(א) ייצוג באמצעות המד"שים המינימליים באוסף (במקרה 1), או על ידי כלל ההופעות כרצף אותיות (במקרה 2).

(ב) ייצוג באמצעות מד"ש *מסויים* של ביטוי הכותרת (במקרה 1), או על ידי הופעה *מסויימת* כרצף אותיות (במקרה 2).

דוגמא א': זו דוגמא למקרה 1(א) דלעיל. ב"מבוא למדגמי כותרת" (סעיף א'), מתואר מדגם כותרת, שהוכן על פי שתי מובאות מקבילות מן התלמודים, הירושלמי והבבלי.

ביטוי הכותרת: "בתשרי".

שאר הביטויים:

א. העולם.
ב. נולדו.
ג. האבות.
ד. נפקדו.
ה. אמהות.

במדגם כותרת יש קשר מושגי בין הכותרת לבין ביטויי גוף המדגם, אך בדרך כלל אין קשר בין הביטויים לבין עצמם. בדוגמא שלפנינו, הכותרת "בתשרי" קשורה באמצעות המובאות הנ"ל עם כל אחד משאר הביטויים. אולם שאר הביטויים אינם קשורים *בהכרח* זה בזה: למשל, אין קשר בין "העולם" ל"נולדו" או ל"נפקדו".

בשלב ראשון, נחשב את המפגשים בין הכותרת לבין כל אחד מן הביטויים האחרים, דהיינו המִפגשים של זוגות הביטויים:

א. "בתשרי – העולם",
ב. "בתשרי – נולדו",
ג. "בתשרי – האבות",
ד. "בתשרי – נפקדו",
ה. "בתשרי – אמהות".

הכותרת אינה מצויה בבראשית כרצף אותיות, ולכן ניתן להשתמש בה כמד"שים בלבד. בכל החישובים, "בתשרי" הוא המלה "הראשונה", והיא נלקחת אך ורק בדילוג השווה, בעוד שההופעות של שאר הביטויים בדילוג השווה, מתחרות עם ההופעות שלהם בדילוגים המשובשים, על המפגשים עם "בתשרי". כך נקבל קבוצה של תוצאות $c(w,w')$, ונחשב עבורן את ערכי P_1 ו- P_2. מטרתנו היא לקבוע אם ערכים אלה "נמוכים באופן חריג".

דוגמא ב': זו דוגמא למקרה 2(א) דלעיל. ב"מבוא למדגמי כותרת" (סעיף ב'), מתואר מדגם כותרת בעניין פקידתה של רחל.

ביטוי הכותרת: "ויזכר א/להים את רחל".

שאר הביטויים:

א. א' תשרי.
ב. בא' תשרי.
ג. א' בתשרי.
ד. נפקדה.
ה. נפקדת.
ו. הנפקדת.
ז. ביוסף.

בשלב ראשון, נחשב את המפגשים בין הכותרת לבין כל אחד מן הביטויים האחרים, דהיינו המִפגשים של זוגות הביטויים:

א. "ויזכר א/להים את רחל – א' תשרי",
ב. "ויזכר א/להים את רחל – בא' תשרי",
ג. "ויזכר א/להים את רחל – א' בתשרי",
ד. "ויזכר א/להים את רחל – נפקדה",
ה. "ויזכר א/להים את רחל – נפקדת",
ו. "ויזכר א/להים את רחל – הנפקדת",
ז. "ויזכר א/להים את רחל – ביוסף".

הכותרת מצויה בבראשית אך ורק כרצף אותיות, ולא כמד"שים. בכל החישובים, הכותרת "ויזכר א/להים את רחל" היא המלה "הראשונה", והיא נלקחת כרצף אותיות, בעוד שההופעות של שאר

הביטויים בדילוג השווה, מתחרות עם ההופעות שלהם בדילוגים המשובשים, על המפגשים עם הכותרת (ראה "מדידת מפגשים" סעיף ג'). כך נקבל קבוצה של תוצאות $c(w,w')$, ונחשב עבורן את ערכי P_1 ו- P_2. מטרתנו היא לקבוע אם ערכים אלה "נמוכים באופן חריג".

דוגמא ג': זו דוגמא למקרה 2(ב) דלעיל. ניצור דוגמא זו מן הדוגמא הקודמת. במקום הכותרת: "ויזכר א/להים את רחל", נבחר ככותרת רק את המלה "ויזכר" (שהיא המלמדת על הפקידה). אולם, המלה "ויזכר" מופיעה בספר בראשית גם בשלושה מקומות אחרים (למשל, לגבי נח). על כן *נבודד ונסמן* את המלה "ויזכר" בביטוי "ויזכר א/להים את רחל": יש כאן רצף של 5 אותיות המתחיל באות שמספרה הסידורי 42,438 מתחילת הספר. להלן נסמן רצף אותיות זה באמצעות השלֶשה (42438, 1, 5), כאשר המספר הראשון משמאל מציין את מיקום האות הראשונה מתחילת ספר בראשית, המספר השני מציין את גודל הדילוג (רצף אותיות הוא כב"דילוג" 1), והמספר השלישי מציין את מספר האותיות בביטוי המסומן. זו תהיה אפוא הכותרת בדוגמתנו.

ביטוי הכותרת: (42438, 1, 5) (= המלה "ויזכר" המסומנת בביטוי "ויזכר א/להים את רחל" שבטקסט).

שאר הביטויים:

א. א' תשרי.
ב. בא' תשרי.
ג. א' בתשרי.
ד. נפקדה.
ה. נפקדת.
ו. הנפקדת.
ז. ביוסף.

בשלב ראשון, נחשב את המפגשים בין הכותרת לבין כל אחד מן הביטויים האחרים, דהיינו המִפגשים של זוגות הביטויים:

א. "(42438, 1, 5) - א' תשרי",
ב. "(42438, 1, 5) - בא' תשרי",
ג. "(42438, 1, 5) - א' בתשרי",
ד. "(42438, 1, 5) - נפקדה",
ה. "(42438, 1, 5) - נפקדת",
ו. "(42438, 1, 5) - הנפקדת",
ז. "(42438, 1, 5) - ביוסף".

ההמשך הוא כמו בדוגמא ב'.

ב. מבחן ראנדומיזציה למִדגמים מטיפוס S (מדגמי "כותרת")

בסעיף זה נתאר פורמאלית את הראנדומיזציה. הקורא שאינו מתעניין בצד הפורמאלי, רשאי לדלג לסעיף ג', שם נתאר את התהליך באמצעות דוגמא מוחשית.

מדגם מטיפוס S הוא מדגם א-סימטרי הניתן להצגה כקבוצת זוגות הביטויים

$$(w,w_1),(w,w_2),\cdots,(w,w_n).$$

בהם הביטוי w ("הכותרת") משותף לכל הזוגות. לגבי מדגמים מסוג זה יתבצע מבחן הרנדומיזציה באופן הבא: לכל $1 \le i \le n$, אותיות הביטוי w_i יעברו פרמוטציה π_i, כך שיתקבל הביטוי $w(\pi_i)$.

נסמן ב- q סדרה של n פרמוטציות π_i. בהפעלת q מתקבל המדגם המשובש:

$$(w, w(\pi_1)), (w, w(\pi_2)), \ldots, (w, w(\pi_n))$$

עבור מדגם זה, מקבלת הסטטיסטיקה j (שהיא מידת "הנטייה הכוללת לקרבה") את הערך P_j^q.

נסמן ב- Q את סך כל סדרות הפרמוטציות q, השונות זו מזו (Q שווה למכפלת הפרמוטציות *השונות* שאפשר לבצע לגבי כל ביטוי w_i, $1 \le i \le n$). באמצעות Q סדרות הפרמוטציות נוצרים Q מדגמים שונים (אחד מהם הוא המדגם המקורי).

את Q ערכי P_j^q המתקבלים עבור Q המדגמים המשובשים, נסדר לפי הסדר הרגיל של המספרים הממשיים. אם התכונה הנמדדת אינה אלא אקראית, הסיכוי ש- P_j (ערך הסטטיסטיקה j עבור המדגם המקורי) יאכלס כל אחד מ- Q המקומות בסדור זה, הינו שווה. זו היא השערת האפס שלנו.

במקרה שהמספר Q גדול מאד, לא נוכל לחשב את כל ה- P_j^q עבור כל Q המידגמים הנ"ל. כדי לחשב את רמת המובהקות הסטטיסטית, נגריל M סדרות q של n פרמוטציות π_i (בדרך שתוארה ב"ראנדומיזציה"). כל אחת מסדרות הפרמוטציות q קובעת את הסטטיסטיקה P_j^q; יחד עם P_j יש לנו $M+1$ מספרים. נגדיר את הדירוג של P_j בתוך $M+1$ המספרים הללו, כמספר ה- P_j^q שאינם קטנים מ-P_j; אם P_j שווה ל- P_j^q אחרים, חציים של אלה ייחשב כ"מקדים" את P_j בדירוג.

נסמן ב- r_j את הדירוג של P_j, מחולק ב- $M+1$. בהשערת האפס, r_j הוא ההסתברות ש-P_j קרוב כל כך לראש הדירוג.

ג. דוגמא לביצוע מבחן ראנדומיזציה למִדגמים מטיפוס S (מדגמי "כותרת")

נדגים כאן את ביצוע מבחן הראנדומיזציה עבור דוגמא ב' שבסעיף א' דלעיל.

1. לוקחים (לפי הסדר) אחד מצמדי הביטויים שבמדגם. עורכים ל"מלה השניה" שבצמד 120 פרמוטציות (כולל סידור האותיות המקורי). במקרה שמספר הפרמוטציות השונות האפשריות, k, קטן מ- 120, מבצעים k פרמוטציות. הפרמוטציות מבוצעות בשיטה קבועה באמצעות תוכנה, שהוכנה בידי יעקב רוזנברג.

לדוגמא: הצמד הרביעי הוא "ויזכר א/להים את רחל – נפקדה". נציג כאן כמה מן הזוגות הנוצרים על ידי הפרמוטציות הנ"ל (כסדרן, החל מן הראשונה – שהיא הסידור המקורי - ועד האחרונה, משמאל לימין):

ויזכר וכו' הנדקפ	ויזכר וכו' הנדפק	...	ויזכר וכו' פנדהק	ויזכר וכו' פדהנק	ויזכר וכו' פדנהק	ויזכר וכו' נפקדה

(למלה "נפקדה" יש 120 פרמוטציות, אך כיוון שאין אנו מבדילים בין מלה בדילוג לפנים, לבין מלה הפוכה בדילוג לאחור – יוצא כי למעשה ישנן 60 פרמוטציות שונות. כלומר, במקרה זה $k = 60$).

2. מחשבים את ערכי $c(w,w')$ של המפגשים בכל אחד מ- 120 (או k) הביטויים המשובשים עם הכותרת. למשל, לגבי הדוגמא דלעיל אנו מקבלים שורה של תאים. בכל תא יש ערך c של הזוג המסויים. תא ריק – פירושו שהפרמוטציה של הביטוי לא הופיעה בדילוגים שווים (בדוגמא שלנו לא אירע מקרה כזה):

72/125	46/125	•••	84/125	120/125	66/125	81/125

3. אם ביטוי במדגם מהווה חלק מביטוי אחר במדגם, יש לוודא שיחס זה ישמר גם בפרמוטציות שלהם.

לדוגמא: תצורת התאריך "א' תשרי" כלולה בתצורה "בא' תשרי". כפרמוטציות של "בא' תשרי" נלקחו הפרמוטציות של "א' תשרי" עם "ב" בראשן. הנה כמה מן הזוגות הנוצרים על ידי הפרמוטציות של "א' תשרי" (כסדרן, החל מן הראשונה – שהיא הסידור המקורי - ועד האחרונה, משמאל לימין):

ויזכר וכו' יארתש	ויזכר וכו' יארשת	...	ויזכר וכו' שארית	ויזכר וכו' שריאת	ויזכר וכו' שראית	ויזכר וכו' א' תשרי

וערכי c שלהם הינם:

85/125	111/125	•••	82/125	81/125	31/125	6/125

ובמקביל, הפרמוטציות עבור "בא' תשרי" נותנות את הזוגות הבאים:

ויזכר וכו' ביארתש	ויזכר וכו' ביארשת	...	ויזכר וכו' בשארית	ויזכר וכו' בשריאת	ויזכר וכו' בשראית	ויזכר וכו' בא' תשרי

ואת ערכי c:

79/125	120/125	•••	47/125	98/125	41/125	10/125

משבצים מספרים אלה בשורת תאים אחת: בכל תא **שני** ערכי c: האחד עבור הפרמוטציה של "א' תשרי" ואחד עבור הפרמוטציה המקבילה של "בא' תשרי":

85/125 79/125	111/125 120/125	•••	82/125 47/125	81/125 98/125	31/125 41/125	6/125 10/125

4. שלבים 1, 2 ו- 3 מבוצעים לגבי כל הצמדים במדגם. כך מקבלים שורות של תאים, בכל שורה 120 (או k) תאים. בכל תא שאינו ריק נמצאים ערכי $c(w,w')$ – אחד, שנים או יותר – של מפגשי "כותרת – ביטוי משובש", כפי שהוסבר ב-1, 2 ו-3.

5. אחר כך מגרילים אחד מן התאים בשורה הראשונה, אחד מן התאים בשורה השניה, וכן הלאה. מקבלים קבוצה של ערכי $c(w,w')$, ומחשבים את ערכי P_i עבורם.

6. חוזרים על התהליך מספר גדול של פעמים, M, באמצעות הגרלות אקראיות שנוצרו באופן דומה לזה המתואר ב"ראנדומיזציה". התוכנה הנדרשת לכך הוכנה אף היא בידי יעקב רוזנברג.