

תוכן עניינים

יא.....	הקדמה
1.....	מבוא
3.....	פרק 0
9.....	פרק 1: ערכים מרכזיים – ממוצע, חציון ושכיח
12.....	ערכים מרכזיים
12.....	הממוצע
12.....	החציון
13.....	השכיח
17.....	הצגה גרפית של נתונים
20.....	יישום
29.....	פרק 2: מדדי פיזור
30.....	מדד לפיזור כמאפיין סטיות
31.....	מדדי פיזור
31.....	ממוצע הערכים המוחלטים של הסטיות מהממוצע
32.....	שונות
32.....	סטיית תקן
34.....	פיזור מבטא מרחק
38.....	נוסחת קיצור לחישוב השונות
38.....	החוק האמפירי
41.....	יישום א
44.....	יישום ב
47.....	יישום ג
53.....	הערכה מהירה של סטיית התקן

54.....	ציון תקן
57.....	שינוי קנה המידה
58.....	הזזה
61.....	יישום ד
63.....	פרק 3 : רגרסיה
65.....	דיאגרמת פיזור
67.....	עקרון הריבועים הפחותים
88.....	אפקט הרגרסיה
90.....	מקדם המתאם [הליניארי]
99.....	Root Mean Square Error (R.M.S.E.)
102.....	מתאם לעומת סיבתיות
103.....	פרק 5 : הסתברות
107.....	הגדרה של מאורעות
111.....	הגדרת ההסתברות
117.....	סידור לפי התור
120.....	התנסות בקומבינטוריקה
122.....	פרק 6 : הסתברות מותנית
127.....	נוסחת ההסתברות המותנית
129.....	חוק המכפלה
130.....	חשיבות של הסתברויות מותנות
133.....	אי-תלות
135.....	חוק המכפלה למאורעות בלתי תלויים
136.....	משפט ביאס
140.....	יישום
143.....	פרק 7 : משתנים מקריים
147.....	פונקציית ההסתברות של משתנה מקרי
150.....	סדרת ניסויי ברנולי והמשתנה המקרי הבינומי
152.....	אפיונים של משתנים מקריים
153.....	התוחלת של משתנה מקרי

157.....	חוקים טכניים של תוחלת
159.....	השוונות של משתנה מקרי
160.....	חוקים טכניים של שונות
163.....	יישום
167.....	נספח
169.....	פרק 8 : ההתפלגות הנורמלית
172.....	טבלת התפלגות נורמלית סטנדרטית
179.....	הממוצע של תצפיות כמשתנה מקרי
182.....	חוק המספרים הגדולים
184.....	משפט הגבול המרכזי
185.....	ציון התקן של ממוצע המדגם
185.....	ציון התקן של השכיחות היחסית
186.....	יישום
189.....	פרק 9 : הסקה סטטיסטית
190.....	בדיקת השערות
193.....	טעות מסוג ראשון
193.....	טעות מסוג שני
196.....	מבחנים סטטיסטים
198.....	אזור הדחייה
198.....	אזור הקבלה
198.....	מבחן סטטיסטי
206.....	הערה טכנית חשובה
207.....	מבחנים חד-צדדיים
210.....	רמת המובהקות α
212.....	פרק 10 : בדיקת השערות – המשך
215.....	מבחן t חד-מדגמי
217.....	לוח התפלגות t
219.....	מבחנים לשני מדגמים בלתי תלויים
223.....	מבחן t למדגמים מזווגים

225.....	השוואה בין שתי פרופורציות
228.....	p-value ה-
229.....	תרשים זרימה לבדיקת השערות
231.....	הסתברות לטעות מסוג שני
232.....	גודל המדגם
234.....	יישום
236.....	פרק 11 : בדיקת השערות לאי-תלות ולהומוגניות באמצעות לוחות שכיחות
239.....	לוח התפלגות חי בריבוע χ^2
244.....	יישום
245.....	פרק 12 : אומדים ורווחי סמך
246.....	הטיה של אומד
247.....	רווחי סמך
248.....	רווח סמך עבור פרופורציה
250.....	רווח סמך וגודל המדגם
251.....	רווח סמך עבור תוחלת μ
252.....	רווחי סמך אחרים
254.....	תרשים זרימה לרווחי סמך
255.....	פרק 13 : יישום אינטגרטיבי
266.....	פרק 14 : אחרית דבר
267.....	נספח 1 : סיגמא Σ
272.....	נספח 2 : מערכת צירים קרטזית
273.....	משוואת הישר במערכת הצירים
274.....	מציאת המשוואה של ישר
275.....	שרטוט קו ישר
277.....	סיפורים
295.....	תרגילים
318.....	פתרונות
383.....	איורים ומקורותיהם



מבוא

הביטוי "סטטיסטיקה" מוצאו ב"status" הלטיני, והוא קשור ל"state" האנגלי: משמעות המילה במקור היתה אוסף של נתונים על מצב המדינה. עם הזמן המקצוע התפתח: הובן שלא מספיק לאסוף נתונים; צריך לדעת גם איך לנתח אותם. אפילו באיסוף נתגלו קשיים: עם גידול האוכלוסייה או התופעה הנחקרת הפך מפקד מלא להיות לא מעשי, ונוצר צורך להסתפק במדגם. כיצד להגיע למדגם מייצג? מה צריך להיות גודלו כדי שיאפשר מסקנות מהימנות? כיצד להסיק ממדגם על אוכלוסייה שלמה?

מתברר שכדי להסיק ממדגם, עליו להיות "מקריי". "מקרה" היא מילה ישנה. (תרגיל מעניין הוא לבדוק בקונקורדנציה את הופעות המילה בתנ"ך, ולנסות לעמוד על הצורה בה הקדמונים הבינו אותה.) העולם העתיק תפס את המציאות במונחים הרבה יותר דטרמיניסטיים מאיתנו, דבר שאולי מנע את התיאור של תופעות אקראיות בצורה כמותית. התורה הכמותית – המודרנית – העוסקת בתופעות מקריות היא תורת ההסתברות, שתחילת פיתוחה במאה ה-17. המניעים לפיתוחה היו שאלות שהתעוררו בהימורים על משחקי קוביה. עברו כ-200 שנה עד שהובן שאת התורה הזאת ניתן להחיל גם על תופעות מקריות אחרות. אי לכך, הסטטיסטיקה כמקצוע מודרני הוא מקצוע צעיר – ההיסטוריה שלו היא בת 150 שנה בלבד.

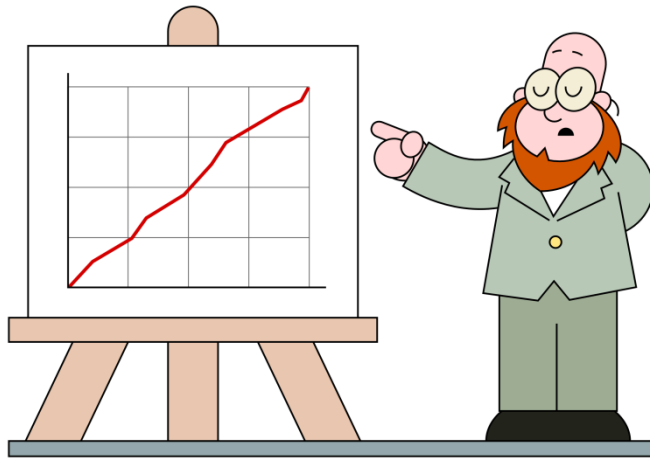
מדוע חייב מדגם להיות "מקריי"? עד כמה שהדבר יראה מפתיע, התשובה היא כי **על המקרה ניתן לסמוך**. אם בסקר דעת קהל נעוט על 50 איש חפים מפשע בצאתם בצהריים מסופרמרקט ירושלמי כדי לבדוק את תגובת הציבור למפלגה "שלנו" בבחירות הבאות, הרי שהמדגם לא ייצג אפילו את הציבור הירושלמי: יהיה ייצוג חסר לשכירים (שבשעת צהריים נמצאים בעבודה),

לסטודנטים (שבשעת צהריים ישנים או נוכחים בהרצאות), לאוכלוסייה שמעדיפה לקנות בשוק, וכו', ויהיה ייצוג יתר לעקרונות בית, לפנסיונרים, וכו'. קשה לבחור באופן מודע מדגם שייצג את כל האוכלוסייה: האם חשבנו על כל האפשרויות? האם דאגנו לייצוג הולם של מעשנים? קשישים? חרדים? חיילים? מורים? חולים? וכך הלאה. באופן פשטני, מדגם מקרי הוא הגרלה הוגנת בין פרטים באוכלוסייה: כאילו לוקחים פתקים ורושמים עליהם את שמות כל הפרטים, מערבבים את הפתקים בכד ומגרילים 50 (או גודל מדגם רצוי כלשהו) מהם. גם מדגם מקרי לא ייצג את כל האפשרויות בדיוק לפי חלקם באוכלוסייה, אבל תורת ההסתברות מלמדת אותנו שעל המקרה ניתן לסמוך, כמו, למשל, בהטלת מטבע הוגן. אמנם נתייחס לתוצאה בהטלה הקרובה כמאורע מקרי, אך "יש סדר בבלגן": למרות שאין ביכולתנו לנבא את התוצאה של ההטלה הבאה, אם נטיל את המטבע 1,000,000 פעמים, נהיה די בטוחים שכ-500,000 מההטלות יסתיימו בצדו האחד של המטבע כלפי מעלה ו-500,000 בצדו האחר. (אם נקבל סטייה גדולה מ-500,000 נתהה אם המטבע הוגן!) באופן דומה, במדגם מקרי מספיק גדול, נצפה שהפרופורציה במדגם של מעשנים תהיה דומה לפרופורציה באוכלוסייה (אף אם אין לנו יודעים את ערכה!), שפרופורציית הקשישים במדגם תדמה לזו שבאוכלוסייה, וכך הלאה. אמנם, הייצוג לא יהיה מדויק, אבל תורת ההסתברות תאפשר לנו לקבל מושג על הטעויות, דבר שלא ניתן לעשות במדגם שאינו מקרי. בסיכומו של דבר, כדי להבין שיטות סטטיסטיות יש צורך בהבנת מושגים בסיסיים בתורת ההסתברות.

הספר מאורגן באופן הבא:

הפרק הטרומ-ראשוני מהווה אשנב לחשיבה סטטיסטית ונועד להראות שהידיעות האינטואיטיביות של הקוראים בנושא עמוקות. ארבעת הפרקים הבאים עוסקים בהצגה של נתונים ואפיונם בעזרת מדדים, ובהדגמת הרעיון העומד ביסודה של כל תחזית סטטיסטית. ארבעת הפרקים הבאים עוסקים בתורת ההסתברות. שאר הפרקים בספר עוסקים בהסקה סטטיסטית. בכל פרק מובאות דוגמאות. מומלץ לחשוב עליהן באופן עצמאי לפני הסתכלות בפתרון. בסוף הספר תרגילים לכל פרק ופרק ולאחריהם פתרונות.

אחרי שראיינתי אלפי אנשים, מצאתי שקיים מתאם חזק בין חכמה לבין הסכמה אתי.



פרק 4

מקדם המתאם [הליניארי] (The [Linear] Correlation Coefficient)

בפרק הקודם למדנו על רגרסיה, מודל לניבוי ערך Y על סמך ערך X באמצעות קו ישר. בדומה לנעשה בפרק 1, שם השתמשנו ב- \bar{X} לבטא ערך טיפוסי של סדרה ולא הסתפקנו בזה, אלא רצינו בפרק 2 להציג מדד לטיב התיאור (ס כמדד לפיזור סביב \bar{X}), גם בנושא רגרסיה יש עניין להגדיר מדד לטיב התיאור של הקשר בין X לבין Y (מדד שיתאר עד כמה טוב הניבוי של Y לפי X).

המדד המקובל הוא **מקדם המתאם הליניארי**. בדרך כלל משמיטים את המילה "הליניארי", ומדברים על "מקדם המתאם", אבל יש לזכור שהכוונה למדד של טיב התיאור על ידי קו ישר. ייתכן שקו שאינו ישר יבטא את הקשר באופן יותר טוב.

בפרק זה נשוב ונדון בדוגמאות מהפרק שעבר ונראה כיצד מקדם המתאם מאשש את התוצאות שקיבלנו בניתוחי הרגרסיה.

מקדם המתאם מסומן בדרך כלל על ידי האות r (ולפעמים על ידי ρ). ההגדרה היא:

$$r = \frac{\sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{[\sum_{i=1}^n (X_i - \bar{X})^2] \times [\sum_{i=1}^n (Y_i - \bar{Y})^2]}}$$

ברגע הראשון לא ברור מדוע דווקא ביטוי זה מבטא את חוזק הקשר בין X ל-Y. האינטואיציה העומדת מאחורי המדד מבוססת על פיתוח מתמטי שעיקרו מובא בסוף הפרק הזה.

דוגמה לשימוש בנוסחה

נמצא את מקדם המתאם r לקשר בין עישון ורמת התמותה בהסתמך על הנתונים מהפרק הקודם. בטבלה שבעמוד 69 מצאנו ש:

$$\sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})] = 330575$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = 1396304$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = 148200$$

לכן מקדם המתאם בין עישון לרמת התמותה הוא:

$$r = \frac{330575}{\sqrt{1396304 \times 148200}} = 0.7267$$

בהמשך נלמד כי מתאם זה הוא מתאם די גבוה ומעיד על קשר די חזק וחיובי בין עישון לתמותה.

למקדם המתאם יש מספר תכונות (וניתן להוכיחן באופן מתמטי):

- (1) מקדם המתאם בין Y לבין X זהה למקדם המתאם בין X לבין Y
 - (2) תמיד מתקיים ש- $-1 \leq r \leq 1$
 - (3) אם $|r| = 1$ אם ורק אם כל הנקודות בדיאגרמת הפיזור נמצאות על קו ישר אחד
 - (4) אם $r > 0$ הקשר בין X לבין Y הוא קשר עולה. אם $r < 0$ הקשר בין X לבין Y הוא קשר יורד
 - (5) אם אין קשר בין X לבין Y, מקדם המתאם יהיה 0 (בקירוב)
 - (6) מקדם המתאם לא ישתנה אם נשנה את קנה המידה של X ו/או של Y
 - (7) מקדם המתאם לא ישתנה אם נוסיף קבוע כלשהו ל-X ו/או ל-Y
- (במילים אחרות: מקדם המתאם בין X לבין Y זהה למקדם המתאם בין $rX + s$ לבין $tY + u$ אם r ו-t קבועים חיוביים כלשהם, ואם s ו-u קבועים כלשהם.)
- (8) מקדם המתאם הוא מספר טהור: אין לו קנה מידה והוא לא תלוי בקנה המידה של המשתנים X, Y